Implicit Learning of Perceptual Comparisons with Extended Training

A thesis submitted in fulfilment of the requirements for the Degree of Master of Science in Psychology

by Timothy Y. Chen

Department of Psychology

University of Canterbury

2019

Abstract

How do we compare stimuli that vary in magnitude? An overview of relevant literature shows that perceptual comparisons have been a longstanding topic of interest in various subfields of psychology, and there currently is a mixed view in regards to how cognitive systems encode and compare stimulus magnitudes. In particular, a conjecture put forth by Torgerson (1961) states observers use a single quantitative relation to compare magnitudes, and that relation is either a difference or ratio. Torgerson's (1961) conjecture has stimulated considerable research using stimuli of various modalities; these studies have typically used experimental designs such as magnitude estimation in which responses are symbolic (i.e., numbers) and observers can use their mathematical knowledge. Recently Grace et al. (2018) developed a novel, nonsymbolic task – an 'artificial algebra' – to study perceptual comparison, and found results inconsistent with Torgerson (1961). Their results showed significant control by both difference and ratio relations over responding, which contradicts Torgerson's (1961) prediction that observers use a single operation.

The research in this thesis used Grace et al.'s (2018) artificial algebra task with the primary aims of replicating and extending their results. We conducted two experiments, with brightness and line length as the stimulus modalities. For each experiment, eight observers were randomly assigned to receive training with difference or ratio relations. They completed four sessions on separate days, so we could assess the effect of extended training. On each trial, observers saw a pair of stimuli and made a response on an analogue bar based on comparing the stimulus magnitudes. Feedback was provided based on the difference or ratio of the nominal stimulus values. Results were similar to those of Grace et al. (2018) and showed that observers responded accurately in both experiments. Although there was relatively little improvement in measures of accuracy over sessions, multiple regressions showed that control by the trained relation increased while control by the untrained relation decreased. This confirms that

feedback provided over multiple sessions was effective in terms of training observers to respond based on either difference or ratio relations. Results also showed that for the majority of observers and sessions, there was significant control by the untrained relation. These results are consistent with those of Grace et al. and show that responding in their task is based on two operations rather than one, contrary to Torgerson's conjecture.

Our perceptual and cognitive systems allow us to estimate the magnitude of a stimulus and make comparisons based on those estimates. These abilities are integral to our general functioning as we move about and interact with objects in our environment, albeit rather unconsciously. Such examples are ubiquitous in daily life; we can quite easily tell how far we need to reach for an object away from us, or which of two lines is longer. Our ability to make perceptual comparisons seems automatic and even effortless, but has raised many interesting questions for research. For example, in a prototypical magnitude estimation experiment, subjects are given explicit instructions to judge and assign numerical estimates to the subjective magnitudes, such as the perceived length of a line (Poulton, 1968). Results of these types of experiments indicate that humans can respond accurately (Parker, Schneider, & Kanow, 1975). Research has shown that these abilities are not limited to humans, as non-human primates are also able to estimate magnitudes (Murofushi, 1997; Tomonaga, 2002). Whilst it is fairly established that we are capable of making perceptual comparisons, there currently is a lack of deeper understanding thereof. How is magnitude represented in human cognition? How are comparisons accomplished? What are the underlying processes when observers compare stimuli magnitude? Further research on perceptual comparisons is important given its vital role in human functioning and daily life.

In the following review, literature pertinent to the research of perceptual comparisons is considered. Since the origins of experimental psychology as a scientific discipline, research has investigated topics related to magnitude estimations, including the well-known controversy regarding the form of the psychophysical law (Fechner vs Stevens; Krueger, 1989). Particular attention will be given to *Torgerson's conjecture*, which posits that observers only perceive a single relation between stimuli, and this relation is either a difference or a ratio (Torgerson, 1961). Several studies have investigated this proposal. However, no conclusive evidence has emerged. Grace et al. (2018) adopted a novel procedure in which non-symbolic representations

of magnitude was used. This was coined 'artificial algebra' because neither numbers nor explicit instruction were used. Their experiment examined one's implicit learning of artificial algebra in a perceptual comparison task, which showed conflicting results. In more recent years, studies in the numerical cognition literature have developed theories which could explain the results of a magnitude comparison task. Finally, the present study was built upon Grace et al.'s (2018) paradigm, with a few new ideas incorporated.

The earliest theories of psychophysics were developed in the 19th century. In this time, physicists in Germany such as Gustav Fechner pioneered experimental psychology and developed psychophysical laws derived from empirical reserach, such as Weber's law (Krueger, 1989; Murray, 1993). Many of these laid the groundwork for contemporary psychological theories. In particular, there is a classic debate in regards to quantifying human perception of physical magnitudes. First, Fechner (1860) devised the Weber-Fechner law (Fechner, 1860/1966) which was built on the presupposition of Weber's law. Weber's law functions on the just noticeable difference (JND) between pairs of stimuli. That is, the smallest difference between the strengths of two stimuli that could be reliably detected. Moreover, Weber's law stated JND is proportional to the intensity of the stimuli, so that JND is a constant proportion of the original threshold size. Assuming JND is constant, the Weber-Fechner law proposed that JNDs can be concatenated to produce a single scale of perceived magnitude. Integration gave the logarithm of physical intensity (*I*) as the psychophysical function. This becomes:

$$P = K log I \tag{1}$$

where *P* is the perceived magnitude, and *K* is a constant.

By contrast, Stevens' power law (Stevens, 1957) refuted the above and stated that the perceived magnitude of a stimulus (*P*) equals a constant (*K*) multiplied by the stimulus intensity (*S*), raised to a power. In mathematical form, this becomes:

$$P = KS^n \quad (2)$$

The parameter *n* is the sensitivity or psychophysical exponent, which varies across stimulus modalities (Stevens & Galanter, 1957). Stevens' formula was supported by studies using psychophysical methods such as magnitude estimation and cross-modality matching (Krantz, 1972). In cross-modality matching studies, subjects make magnitude estimations and adjust to the level of modality A to match various levels of modality B by alternating from one modality to another from trial to trial, under instruction to estimate all presentations on a common scale of sensory magnitude (Stevens & Marks, 1980). Such studies have interspersed sessions with combinations of modalities such as brightness and loudness (Stevens & Marks, 1965), loudness, vibration, and electric shock (Stevens, 1959), and area and loudness (Baird, Green, & Luce, 1980). Results from these experiments conformed better to power functions as opposed to Fechner's logarithmic formula.

Considerable discussion has stemmed since Stevens' (1957) rebuttal of the Weber-Fechner law and follow-up studies have attempted to attest to the predictions of the Weber-Fechner law and Stevens' power law. These studies have used methods such as: 1. magnitude estimations, by exploring subjects' sensitivity to dissimilarity between stimuli on a range of modalities such as taste of sodium chloride after adaptation to sodium chloride (Mcburney, 1966), numerosity (Krueger, 1984), and differing accents (Brennan, Ryan, & Dawson, 1975), 2. category scaling, by assigning unique stimuli-specific category rating for different stimuli such as length of lines, (Eisler, 1963), loudness (Garner, 1954), and taste (McBride, 1983), and 3. fractionation, by separating observers into random groups and assigning each group with varying degrees of stimuli strength to see if predictions can be consolidated over a wider stimuli range (Garner, 1954; Laming, 1991). The controversy between Fechner and Stevens has not been fully resolved, despite attempts to reach compromise (Krueger, 1989; Teghtsoonian, 1971).

Regardless, the controversy between these two basic laws of psychophysics was central in later research on perceptual comparisons.

Garner's (1954) research on category scaling for loudness served as the basis for Torgerson (1961). Garner (1954) had subjects produce tones that were in the middle of a louder and quieter tone by equating ratios or differences. Then he developed a loudness scale which predicts perceived loudness in relation to actual loudness based on his discovery that both the ratio and difference group produced the same sound. Torgerson (1961) then postulated observers may only perceive a single quantitative relation between stimuli. That is, observers use a single operation (a difference or a ratio) to compare stimulus magnitudes. This proposal was termed Torgerson's conjecture by Birnbaum and Veit (1974), and has been the focus of considerable research. Studies have found observers can produce numerical estimates of differences and/or ratios using various types of stimulus modalities. These modalities included heaviness (Birnbaum & Veit, 1974; Mellers, Davis, & Birnbaum, 1984; Rule, Curtis & Mullin, 1981), line length (Masin, 2012, 2014; Parker, Schneider, & Kanow, 1975), grayness (Veit, 1978), loudness (Schneider et al., 1976), pitch (Elmasian & Birnbaum, 1984), sweetness (De Graaf & Frijters, 1988) and remembered distance (Birnbaum, Anderson & Hyman, 1989). Depending on the stimulus and instructions, most of these studies have determined that differences and ratios are based on a single comparison operation, thus in line with Torgerson's conjecture. Additionally, a majority of these studies have pointed towards a single difference operation. However, some have showed evidence of dual operations – that is, observers are capable of responding based on both differences and ratios (Rule et al., 1981). Overall, the results of studies on Torgerson's conjecture are mixed and do not present a clear picture in terms of whether observers can only use a single operation when comparing stimulus magnitudes.

Prior studies on Torgerson's conjecture have used explicit instructions and numerical estimates of stimuli. More specifically, they have used tasks that are 'cognitively penetrable' (Zeimbekis & Raftopoulos, 2015), because participants that can estimate magnitudes objectively will be able to assign numbers based on differences or ratios depending on the instructions, using their mathematical knowledge. Moreover, Birnbaum et al. (1989) found participants were able to judge both differences and ratios of distance between pairs of cities.

Grace et al. (2018) argued that since our perceptual system has evolved prior to the invention of mathematics, it is important to test Torgerson's conjecture under conditions that excludes the use of explicit cognition of mathematic knowledge by the observers. Grace et al. (2018) was the first to tackle Torgerson's conjecture using nonsymbolic stimuli. A range of modalities were tested in their experiments, including circle pairs that varied in brightness (1a), number of dots (1b), and area (1c). Upon each trial, participants observed a pair of stimuli sampled from a set of 28 possible pairs, and made a response by clicking along an unmarked horizontal bar below the stimuli. The far left side of the horizontal bar would indicate maximum similarity, whereas the right end of the horizontal bar indicated maximum dissimilarity. Each session consisted of four blocks of 84 trials (336 total), and there was a 2 second inter-trial interval. Each of the 28 stimulus pairs was presented three times in random order for each block. After each response, feedback was provided in the form of green (for correct) or red (for incorrect) ovals, centred on where the observers responded. 'Correct' responses were defined as those within 7% in either direction of the trained value. Incorrect responses induced a correction trial, where observers were able to attempt the same stimulus pair once more. Correction trials reinforced learning of the task, and were omitted from analyses.

The key aspect of Grace et al.'s procedure was that feedback was based on either the difference or ratio of the nominal stimulus values, scaled linearly to the response bar such that the left and right ends meet the minimum and maximum difference (or ratio) across the stimulus

pairs. By capitalizing on feedback training and analogue response, Grace et al. (2018) studied perceptual comparisons without the use of numbers or explicit cognition, and they described their task as an *artificial algebra*, because observers learned arithmetic relations between stimulus magnitudes.

In Grace et al.'s (2018) Experiment 1, observers completed two conditions in which the feedback provided was based on the differences or ratios of the stimulus pairs, in counterbalanced order. In Experiment 1a, observers completed two sessions in each conditions which were run on the same day (morning and afternoon), whereas in Experiment 1b and 1c observers completed one session per condition and were completed on different days. After the first condition was completed, participants were told that how correct responses were determined would change in the next session. They were not told the nature of the change and were instructed to learn to respond accurately based on the feedback presented. Participants were randomly assigned to a difference or ratio group, and the order was counterbalanced.

Torgerson's conjecture would predict better performance when feedback corresponded to the comparison operation used by the perceptual system, either difference or ratio. However, Grace et al. (2018) found that observers learned the task quickly and to about the same level of accuracy in both difference and ratio conditions. Their results showed strong correlations with average individual correlations of r = .95 (difference group) and .94 (ratio group) across experiments between the trained value and average response location in each experiment. The percentage of variance accounted for by regressions of responses on trained values ranged from 93.6% (1b, ratio condition) to 97.9% (1c, difference condition). There was no statistically significant ($p < .05$) between groups trained with difference or ratio relations. The strong fits indicated observers were able to learn and respond accurately to the difference and ratio relations equally well. Average results were representative of individual data.

Furthermore, Grace et al. found approximately half of individual cases showed joint control by both operations, regardless of which relation was trained. This result was confirmed by hierarchical multiple regression – trained values were entered at the first step, and the correct values for the untrained relation were entered at the second step. Although observers were trained on a single relation, the untrained relation also accounted for significant incremental variance beyond the trained relation, and coefficients (beta weights) were positive. Additional analyses based on rank-order pairwise comparisons, non-metric scaling, and psychophysical modelling suggested that observers responded based on two operations rather than one. Importantly, evidence for control by both relations was found in the first condition, so that joint control by both relations could not be attributed to a carryover effect.

A possible limitation that was addressed in Grace et al's data was that the difference and ratio training values were highly correlated (r = .91), which presents potential for multicollinearity problems in the multiple regression analyses. Monte Carlo simulations were used to test the strength of evidence of the regression results for control by two operations. Simulated responses were generated as a linear function of trained difference or ratio values, with added Gaussian noise (σ). Defined weights were given for the difference and ratio training values. A set of simulated response for the 28 stimulus pairs was generated for each of 1000 iterations. Multiple regression models with difference and ratio values as predictors were then fitted to the set, and the estimated parameters were saved. The defined weights and noise parameter were varied across simulations. The Monte Carlo simulations showed the recovered difference and ratio coefficients were nearly identical to the defined weights when the noise was small (σ = .05), and decreased only slightly when noise increased (up to σ = .2). Therefore, whether responses were determined singly or jointly by differences and ratios was identifiable from the regression coefficients. These results suggest that results of multiple regressions from

Grace et al's experiments provided valid information about the relative control by differences and ratios over responding, and that the results were not a by-product of multicollinearity.

Another possible confound that was investigated was the presence of exemplar learning. That is, observers were actually responding to the difference or ratio relation between stimulus pairs, rather than remembering specific instances of stimulus-response pairs. This was examined in Grace et al's second set of experiments (2a-2b) by introducing novel stimulus pairs later in the session and omitting their feedback. Overall, the results of Experiments 2a-2b are similar to those of Experiment 1. Averaged across observers, the correlations between responses and trained values were *M*s = .94 and .93 (2a) and *M*s = .96 and .93 (2b) for the difference and ratio group, respectively. Responses on the new trials were close to values predicted by the trained relation, *r*s = .98 and .94 for Experiments 2a and 2b, respectively. This showed the task successfully captured learning of the difference or ratio relation over remembering specific stimulus-response pairs.

Results of Grace et al's (2018) artificial algebra suggested observers were able to learn and produce both differences and ratios without explicit instruction for pairs of visual stimuli that varied in brightness, numerosity, and area. A substantial number of individual cases showed dual control by both operations even though they were trained on a single operation. The implication of these results is that the perceptual system automatically computes both differences and ratios when comparing stimulus magnitudes, and thus Torgerson's conjecture was false. However, questions remain on how and why Grace et al's artificial algebra generated evidence that differed from the majority of the studies. One clear distinction between Grace et al. (2018) and the previous studies was that the latter have relied on responses using explicit numerical cognition. Specifically, in those studies observers responded by giving numerical estimates, whereas Grace et al. (2018) adopted an analogue response mechanism. To address

these differing outcomes, contemporary research on numerical cognition has shed some light onto understanding response in magnitude comparison paradigms.

These topics are well summarised by Walsh (2003), who brought together a theory of magnitude (ATOM). Pertaining to our interests in perceptual comparisons, ATOM can be described in the following way: 1. there is a common basis between space, time, and quantity whose interconnections form a single general magnitude system, 2. this general magnitude system operates from birth, and 3. the parietal cortex in both hemispheres of the brain is a locus of analogical quantity representation. Walsh's proposal integrates empirical results from a range of studies. The following entails a more detailed review of the relevant studies associated to ATOM and our current understanding of perceptual comparisons is discussed.

Advances in numerical cognition have suggested that humans possess innate numerical ability. Studies have presented infants with two instances of the same stimuli simultaneously with equal strength in numerosity. Thereafter, both instances would display a new representation of the stimuli; one alternative would remain equal in numerosity, whereas the other had changed. Experiments have adopted both visual (number of dots, Berch, 2005) and auditory (acoustic beeps, Lipton & Spelke, 2003) cues. Infants showed they would be more attuned to the numerically changing alternative (Starr, Libertus, & Brannon, 2013). This natural affinity of numbers in infants is referred to as *number sense* (Dehaene, 2001), and supports human capacity for math understanding (Starr, Libertus, & Brannon, 2013). The term number sense has been used to encompass a wide array of mathematically relevant constructs and is built upon the presumption of a general magnitude system (Walsh, 2003). One aspect of number sense, estimation, is associated with quantifying and representing number magnitudes and is directly tied to our interests in perceptual comparisons. Our natural ability to perform magnitude estimations could have stemmed from the development of preverbal number sense,

although the extent to which this association varies depends on the administered estimation task (Tosto et al., 2017).

Number sense features two core systems for numerical quantification, which are the approximate number system (ANS) and the object tracking system (OTS; Piazza, 2011). The ANS is cited as a general cognitive system that supports the estimation of the magnitude of a group without explicit language or symbols. The ANS allows representation of numbers in a nonsymbolic format (for a review, see Feigenson, Dehaene, & Spelke, 2004), and it is involved in the process of estimation (Booth & Siegler, 2006). A defining feature of the ANS is that it represents number in an approximate and compressed fashion, so that pairs of stimuli can be discriminated only if they differ by a given numerical ratio, according to Weber's law (VanMarle, 2015). On the other hand, the OTS is a mechanism by which stimuli are represented as distinct individuals and can be tracked through time and space (Piazza, 2011). To elaborate, the OTS facilitates a mental capacity to individuate and automatically realize identities of specific individuals in a scene and is the source of the implicit principles that guide the acquisition of cognisant counting (Gallistel & Gelman, 1992). One of OTS's defining properties is that it is limited in capacity. Various studies have showed simple features such as colour or orientation of objects can be accurately retained in memory, but that storage is limited to a small number of objects (Alverez & Cavanagh, 2004; Pylyshyn, 2001). Moreover, Alverez and Cavanagh (2004) stated there is an inverse relation between the information load per object and the number of objects that can be stored. In their experiments, participants were able to remember more of simple objects, such as coloured squares, than of other more complex objects, such as Chinese characters or random polygons. The OTS is also applicable to enumeration tasks: subjects were able to determine numerosity in small collections of three or four items with very high accuracy and high speed, even in conditions of very briefly presented or masked stimuli (Pylyshyn, 2001; Revkin et al., 2008). However, for sets with more than

three or four items, enumeration is only possible either via explicit counting, or via approximate estimation, which falls under the computational constraints of the ANS, thus reflecting Weber law's through scalar variability (Gallistel & Gelman, 1992). It would be interesting to investigate which of the ANS or OTS is better suited in explaining the data of perceptual comparisons, as this may provide insight on how perceptual comparisons is done. This can be analysed by looking at the variability of responses among observers; if the variability is scalar, then the ANS is accepted over the OTS. As perceptual comparisons require participants to judge and compare magnitudes of two objects, it is hypothesized the OTS will provide a better fit to the results.

In cognitive neuroscience, researchers have deployed fMRI techniques and found neurons tuned to number (Cohen & Dehaene, 1996; Piazza et al., 2004). Lesion studies of the single units, or 'numerons', suggest that the cognitive representation of number may be closely linked to magnitudes (Walsh, 2003). Experiments have included both humans (adults and four-year-olds) and non-human primates as observers (Nieder & Dehaene, 2009). They found commonality in the fMRI results. Specifically, neurons in the intraparietal sulcus and prefrontal cortex were activated when observers were subjected to number tasks. Zonal firing of neurons suggested a number system responsible for processing numeric information. The connections between the parietal and prefrontal cortex have been shown to be the basis of several functions that rely on estimating magnitudes such as spatial computations coded in action coordinates (Stein, 1992), numerosity estimation, temporal judgements, (Critchley, 1953) and size judgements (Walsh, 2003).

Meta-analysis of neuro imaging data suggested estimation involves both nonsymbolic and symbolic estimation (Cohen Kadosh, Lammertyn, & Izard, 2008). Previous studies on Torgerson's conjecture have relied on symbolic presentations of numbers. For example, set sizes were manifested as number of dots which were explicitly countable (Birnbaum, 1978). In

contrast, nonsymbolic estimation involves nonverbal processing of quantities and numerosities without using numerals (Tosto et al., 2017). Nonsymbolic numerical cognition is limited to approximate quantity representations and rudimentary arithmetic operations (Nieder & Dehaene, 2009). Studies have shown non-human animals, human infants, human children, and adults with no school-based instruction in arithmetic can add and subtract large numbers of visual forms or event sequences (Barth, Beckmann, & Spelke, 2008). When comparing two numbers, responding was faster and more accurate with increasing numerical disparity and decreasing numerical size (Brannon, Wusthoff, Gallistel, & Gibbon, 2001). There is potential dichotomy between symbolic and nonsymbolic number representation during perceptual comparisons reflecting conscious/learned mathematical ability versus innate number sense. Preverbal number sense increases in precision over development, prior to the emergence of language or symbolic counting (Lipton & Spelke, 2003). Gershman and Niv (2013) suggested that perceptual estimation obeys Occam's razor in that humans prioritise simpler explanations of sensory inputs and the enumeration process underlying the subject's performance was not counting but estimation. Therefore, it is arguably more fundamental to investigate perceptual comparisons where learned mathematical ability is excluded. That is, without the use of numbers or explicit numerical cognition. Gibson (1969) noted the importance of understanding both the impact of learning and innate mechanisms. Historically, researchers have focused on perceptual learning processes rather than basic perceptual capacities. In recent years, understanding of how magnitude is processed has been far-reaching. However, these theories have mainly focused on numerosity specifically, so it is difficult to extrapolate these results to other forms of dimensions.

Grace et al. discovered observers learned to respond to the trained relation rapidly, and they showed improvement over blocks of trials. Kahneman's seminal book 'Thinking, Fast and Slow' (2011) suggested the brain contains multiple cognitive systems which can be classified

into two modes of thought. The “type 1” system is agile, quick, instinctive, which taps into innate number sense, whereas the “type 2” system is slower, deliberative, logical, and comprises explicit mental cognition. The type 1 system is referred to as automaticity, and may explain Grace et al.’s results in their task.

As previously noted, observers in Grace et al’s (2018) task were able to show learning of the trained relation without much effort. A well-established phenomenon is that practice leads to improvement in performance (James, 1890). Perceptual learning is used to describe training perception skills such as ability to discriminate two musical tones from one another. Zarate, Riston, and Poeppel (2012) compared musicians and non-musicians with pitch discrimination tasks to investigate the effects of musical expertise on interval discrimination. Musicians showed greater accuracy, faster reaction times, and less variation in accuracy than non-musicians. In other words, JNS decreases dramatically with practice, and this improvement is at least partially retained for subsequent days. However, the interval-discrimination thresholds of both musicians and non-musicians were found only at 100 cents and higher. Extensive musical training appeared to reduce the interval-discrimination threshold down to a musically meaningful interval of a semitone (100 cents), but failed on even smaller magnitudes. They argued a semitone may serve as an intervallic limit to acoustic processing, suggesting the benefits of training can reach asymptote. Another way in which people can learn is through reinforcement conditioning. This involves learning through reinforcement or punishment. Grace et al’s (2018) feedback design induced positive reinforcement which may explain the improved performance over blocks. Observers in Grace et al.’s study each completed one session, and individual cases showed increase in correlations between response and trained relation. Observers were quick to learn the underlying trained relation and improved over blocks, but to what extent responding observers may improve with further training is not well understood.

Grace et al.'s (2018) results have potentially far-reaching implications. A large portion of many living organism's behaviour repertoire more or less requires judgement of differences and ratios. For example, computationally complex behaviour such as spatial navigation requires mental awareness of both differences (distance and time from point A to point B) and ratios (velocity; displacement over time). Understanding of how our cognitive systems estimate and compare magnitudes would provide insight on how such behaviour is accomplished.

The present study was built around the original design by Grace et al. (2018) using implicit, nonsymbolic stimuli. Our main goal was to replicate the results of Grace et al. (2018), and revisit the question of Torgerson's conjecture. Two experiments were presented, differing in modality used. Experiment 1 used brightness as the stimuli. In Experiment 2, a new modality (line lengths) was used. This was placed to seek out whether our results are consistent under our experimental conditions, and the design is applicable to a wider range of modalities. As opposed to set pairs, stimulus values were randomly generated. This was to avoid stimulus-response exemplar learning as different pairs of stimuli were presented across trials in each session. According to Kahneman (2011) and the numerical cognition literature, we hypothesize observers will show rapid learning of the trained relation. Grace et al. (2018) showed improvement of performance over blocks. However, a semi-longitudinal study has not been explored under these experimental conditions. In both our experiments, observers completed four sessions, conducted on separate days. We wanted to test whether performance in our task can improve over sessions. Reinforcement and perceptual learning suggest the trained relation will be stronger relative to the untrained relation, but also performance should improve over extended sessions. We also planned to examine patterns of variability in the data to determine if responding was more consistent with the ANS (which would predict an increasing linear trend) or OTS (which would predict constant variability).

Experiment 1

The primary goal of Experiment 1 was to replicate and extend Grace et al.'s (2018) study, using stimuli that varied in brightness. Observers completed four sessions, on separate days, so that we could test the effects of more extended training in this procedure. Observers were randomly divided into difference (n = 4) and ratio (n = 4) groups, depending on the relation used to generate feedback. Rather than using a limited set of exemplars, stimuli were generated pseudo-randomly so that a different pair of magnitudes was presented on each trial. We planned to test whether measures of accuracy and relative control of responding by differences and ratios changed over sessions, and if patterns of response variability were more consistent with the ANS or OTS.

Method

Participants. A total of 8 adults served as participants (6m, 2f). Observers Db, Ja, Jo, and Ma were randomly allocated to the difference condition, and Ar, K, L, and Wa to the ratio condition. All were naïve to the purpose of the research, and received a NZ$15 voucher for each session completed. All reported normal or corrected-to-normal vision.

Materials. A custom Javascript software run on a Lenovo Thinkpad T520 laptop computer was used to display stimuli and record responses. The laptop had a 15.6" (39.62 cm diagonal) LCD display screen, with a resolution of 1600 x 900 pixels, a maximum brightness

of 231 cd/m^2, and a contrast ratio of 670:1. The screen settings were unaltered throughout all sessions.

Pairs of circles in varying grayscale, each 6 cm in diameter, were used as stimuli in Experiment 1. They were displayed side-by-side, 8 cm from the top of the screen and separated by a 3.3 cm gap. The circles were identical except in brightness. An 8-bit monochrome palette ranged from 0 - 255 (0 = minimum luminance, 0.33 cd/m^2; 255 = maximum luminance, 231 cd/m^2) simulated the greyscale values for the circles. The brightness values were randomly sampled for each trial between a minimum of 50 and maximum of 205, with the lower value always displayed on the left side of the screen. A horizontal grey response bar which measured 16.5 cm in length and .2 cm in width was located 6.5 cm below the stimuli and contained no markings. All items were displayed in front of a black background.

Differences and ratios were calculated for each stimulus pair based on the nominal brightness values. Nominal values were linearly scaled along the response bar from left to right, so that 0% (left) corresponded to the minimum difference (or ratio) and 100% (right) to the maximum difference (or ratio). Feedback was provided for each response based on the scaled difference or ratio value. Correct responses were defined as those within ± 7% of the trained value. Red and green ovals centred on the trained value were used as feedback for incorrect and correct responses, respectively.

Procedure. Participants were tested individually in a quiet room. At the beginning of each session, observers were seated with the laptop and presented with the following instructions on the computer screen: ”In this experiment, you will see a series of pairs of gray circles on the screen. For each pair, you need to compare the brightness of the circles by clicking on the horizontal bar. The purpose is to learn how to compare the brightnesses as accurately as you can. You will receive feedback for each correct response, but if your response

is incorrect, you will have a chance to repeat with the same pair of circles. You should respond at whatever pace feels comfortable and natural for you, and will have several chances to take a break during the session." Each participant completed four sessions, which were conducted on separate days (usually consecutive). Each session contained four blocks of 84 trials, giving 336 trials per session. A break interval was given in between each block.

For each trial, observers clicked along the response bar using the mouse. A black dot with a white circular outline 10 mm in diameter appeared immediately on the response bar marking the location of the response. Following a 100 ms delay, feedback showing the trained value was presented. If the response was correct, a green oval centred on the trained value on the response bar persisted for 500 ms. After feedback, all items on the screen were removed and the next trial was presented following a 2 second inter-trial interval. Incorrect responses produced a red oval centred on the trained value as feedback, followed by a single correction trial in which the same stimuli were repeated. Correction trials are only given once; they are not repeated if the correction trial was failed. Responses on correction trials were omitted for analyses. The task was self-paced; there was no time limit.

Results

Accuracy of performance was assessed with four measures: Average percent correct (i.e., the percentage of responses that produced green feedback ovals), average absolute deviation of responses from correct values, root-mean-square (RMS; i.e., the square root of the average squared deviation of responses from correct values), and correlation of responses with correct values (r). These measures were obtained for each session and participant.

Results are shown in Figure 1 for both difference and ratio groups. Overall, responding became more accurate across sessions, and there were no significant differences between groups. These observations were confirmed by a series of repeated-measures ANOVAs with session as a within-subjects factor and group as a between-subjects factor.

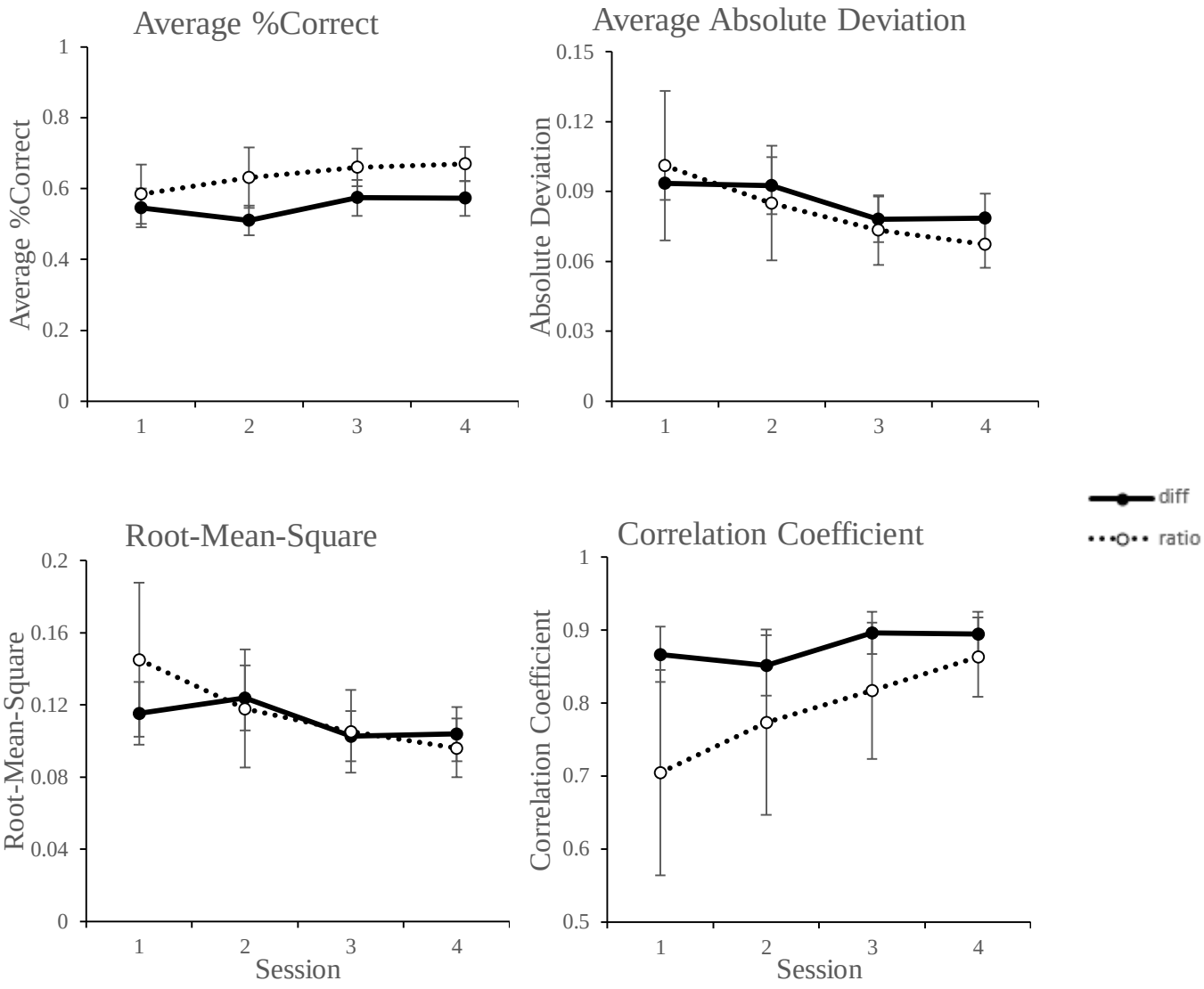

Figure 1. Four measures of accuracy for Experiment 1, average %correct, average absolute deviation, RMS, and R. The solid line represents the difference group (N=4), whereas the dotted line represents the ratio group (N=4).

For average percent correct, the effect of session was significant, $F(3, 18) = 4.535$, $p = .015$, $\eta_p^2 = .430$, but the effects of group and the interaction were not (ps > .244). Post-hoc tests (Newman-Keuls) confirmed that the average percent correct trials for Sessions 1 and 2 were significantly lower than for Sessions 3 and 4. Similar results were obtained with average absolute deviation and RMS. For average absolute deviation, the effect of session was significant, $F(3, 18) = 4.277$, $p = .019$, $\eta_p^2 = .416$, while the effect of group and the interaction were not (ps > .658). For RMS, the decrease across sessions was significant, $F(3, 18) = 4.685$, $p = .014$, $\eta_p^2 = .438$; whereas the effects of group and interaction were not (ps > . 199). For both average absolute deviation and RMS, post-hoc tests (Newman-Keuls) showed that average values for Session 1 were significantly greater than for Sessions 3 and 4.

Correlations of responses with correct values significantly increased across sessions, $F(3, 18) = 4.017$, $p = .024$, $\eta_p^2 = .401$, but again, the effects of group and the interaction were not (ps > .210). Post-hoc tests (Newman-Keuls) confirmed that the average correlation was significantly greater for Session 4 compared to Session 1. Overall, these results suggest that responding became gradually more accurate across sessions, and there were no significant differences in accuracy depending on whether observers were trained with difference or ratio relations. Table 1 reinforces these findings; it shows the averaged means and standard deviations of sessions 1 and 4 of the four measures. The '% improved' column shows the percentage of improvement of each these four measures from the first to the fourth session. All four measures in both groups showed improvement in performance after training, as indicated by an increase in percentage correct and r, and a decrease in absolute deviation and RMS.

Table 1. Results of four measures of accuracy of Experiment 1 (brightness) difference and ratio group. Listed for each measure are the averaged means and standard deviations for the trained relation in sessions 1 and 4. The % improved represents the percentage increase for percentage correct and r, and percentage decrease for absolute deviation and RMS.

Difference Group					Ratio Group				
Measure	Session	Mean	SD	% improved	Measure	Session	Mean	SD	% improved
%Corr	1	0.545	0.111	4.911	%Corr	1	0.584	0.167	14.650
	4	0.572	0.097			4	0.670	0.097	
AbsDev	1	0.093	0.014	16.062	AbsDev	1	0.101	0.064	33.336
	4	0.078	0.021			4	0.067	0.020	
RMS	1	0.115	0.035	10.021	RMS	1	0.145	0.085	33.734
	4	0.104	0.030			4	0.096	0.033	
r	1	0.867	0.075	3.249	r	1	0.705	0.282	22.452
	4	0.895	0.060			4	0.863	0.109	

Figures 2 & 3 shows representative scatterplots of the responses plotted as a function of the scaled trained values. They are 3x4 collated figures showing the difference and ratio group. The rows depict each of the four sessions' data. The left and centre columns show the best and worst performing observer, respectively, in terms of average correlation between response and trained value. Individual session scatterplots were generated by grouping data together in bins based on sorting trials with trained correct values from .00 to .99 in bins with a width of .01. The average response and trained value were then calculated for each bin. The scatterplots' y- and x-axis correspond to these binned averages. The right columns show the averaged result of the four observers in each condition. The mean of each observers' responses and correct values for each bin were used for the averaged scatterplots. The dotted lines show the lines of best fit and the solid line represents perfect correlation. Finally, heterogeneity of variance resulted in clusters at the bottom left quadrants of the scatterplots. As stimuli were randomly sampled from a predetermined range, stimuli pairs were probabilistically more likely to be similar in brightness as opposed to being far apart. In other words, observers were presented with more stimulus pairs which had the correct response on the left side of the response bar (similar) compared to the right side (dissimilar).

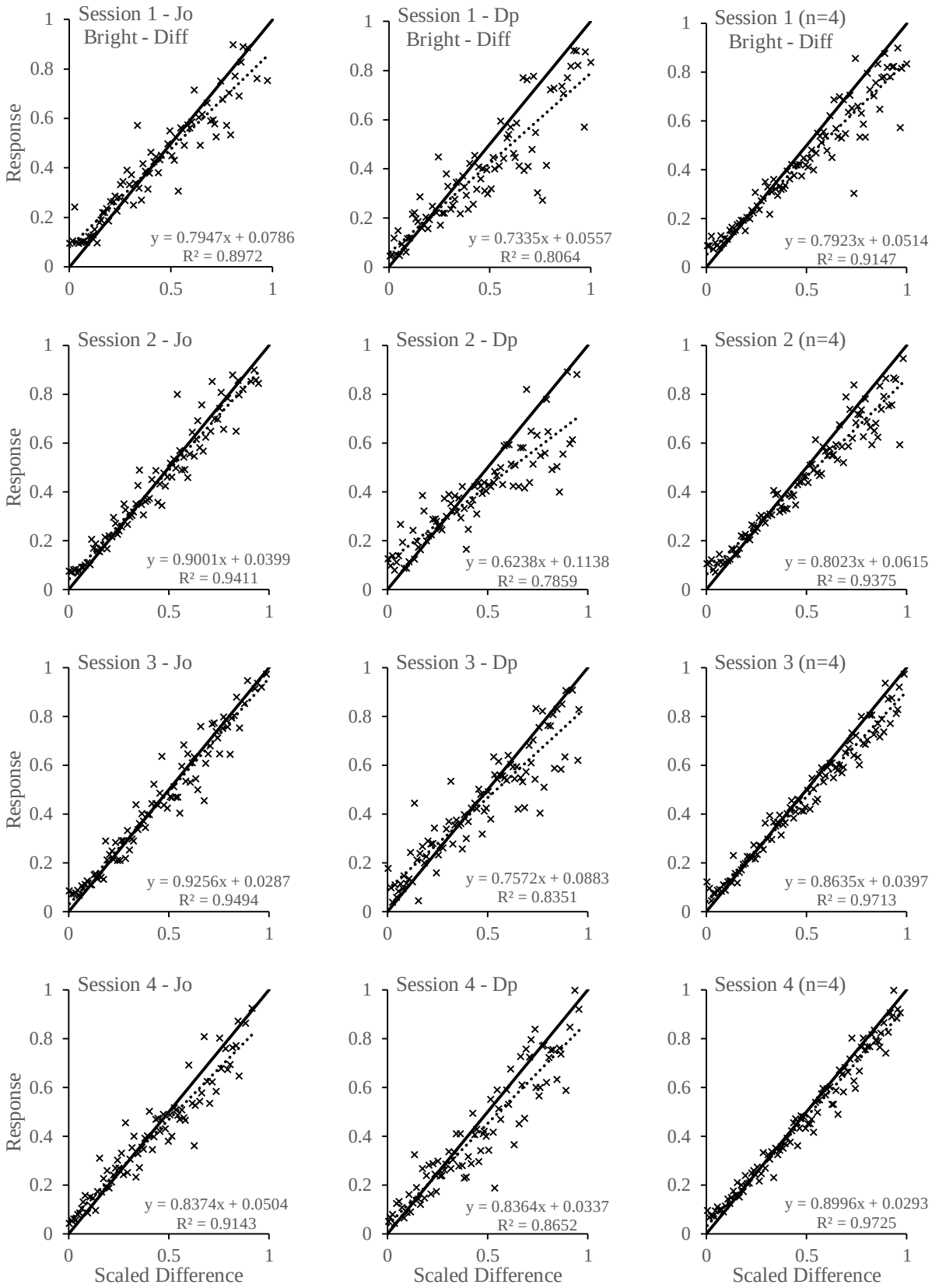

Figure 2. Scatterplots of Experiment 1 difference group as a measure of accuracy. The solid line indicates perfect correlation; the dotted line indicates line of best fit. From left to right shows the best and worst performer, and the averaged results, respectively.

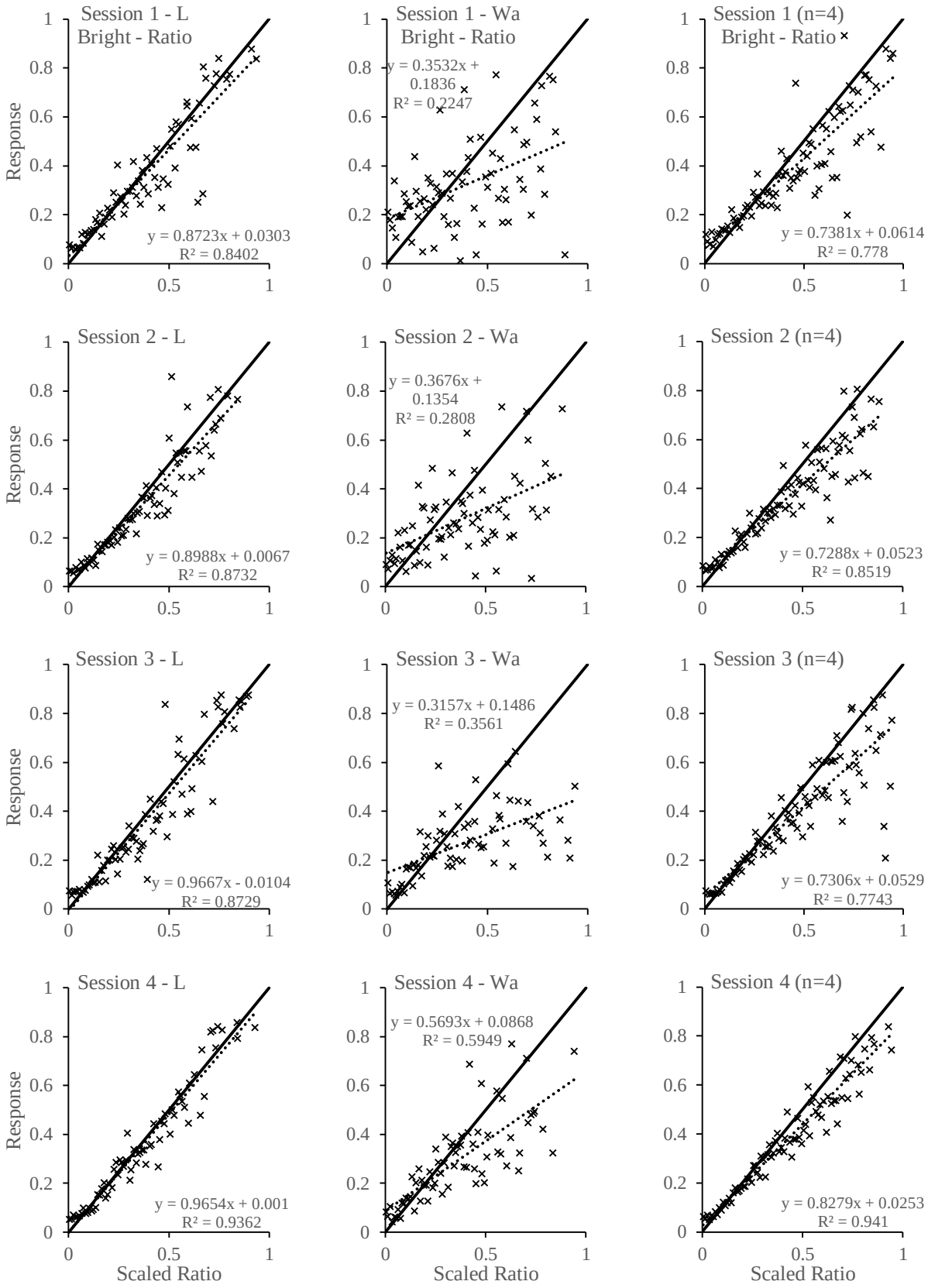

Figure 3. Scatterplots of Experiment 1 ratio group as a measure of accuracy. The solid line indicates perfect correlation; the dotted line indicates line of best fit. From left to right shows the best and worst performer, and the averaged results, respectively.

All observers except one (Wa; Figure 3, centre column) showed a strong positive correlation between responses and correct values. Results for Observer Wa are unusual and will be further discussed later. For the remaining observers, correlations were strongly positive (> .7) in each session. Given no explicit instructions, observers were able to quickly learn and make relatively accurate estimates via simple feedback. Furthermore, averaged correlations improved to well over .9 in the fourth session (Figures 2 & 3).

An important question posed was the degree of control by the two relations, and whether responding was consistent with control by one or two comparison operations. To quantify this, multiple regression was used to determine the degree of control that differences and ratios of brightness values had over responding. Unstandardized regression coefficients (B) gauged the correlations between correct difference/ratio values and response. Thus, they provide comparison of control of participants' responses by the trained and untrained relations. Figure 4 shows the averaged multiple regression trends of the two groups. A prominent feature shown is that the ratio group produced far more significant error bars. The bottom right panel shows the ratio group data presented with the exclusion of observer Wa. With the exclusion of Wa, the significant error bars dissipated. Figure 4 is generally representative of individual data. A repeated measures ANOVA with session and relation (trained vs untrained) as within subjects factors, and group (difference vs ratio) as between subjects factor was conducted. Session was a statistically significant factor, $F(3, 18) = 3.171$, $p = .050$. Also, the interaction between session and relation was found to be statistically significant, $F(3, 18) = 4.950$, $p = .011$. There was no effect of relation. The interaction effect suggested that control by the trained relation increased, whereby control by the untrained relation decreased over sessions. Figure 5 depicts the individual multiple regression analysis of the 8 observers. Over sessions, the B weights for the trained relation increased for all observers. Moreover, the opposite trend is

found in the untrained relation whereby their B weights decreased over sessions. Combining our results thus far, observers were successful in learning the trained relation.

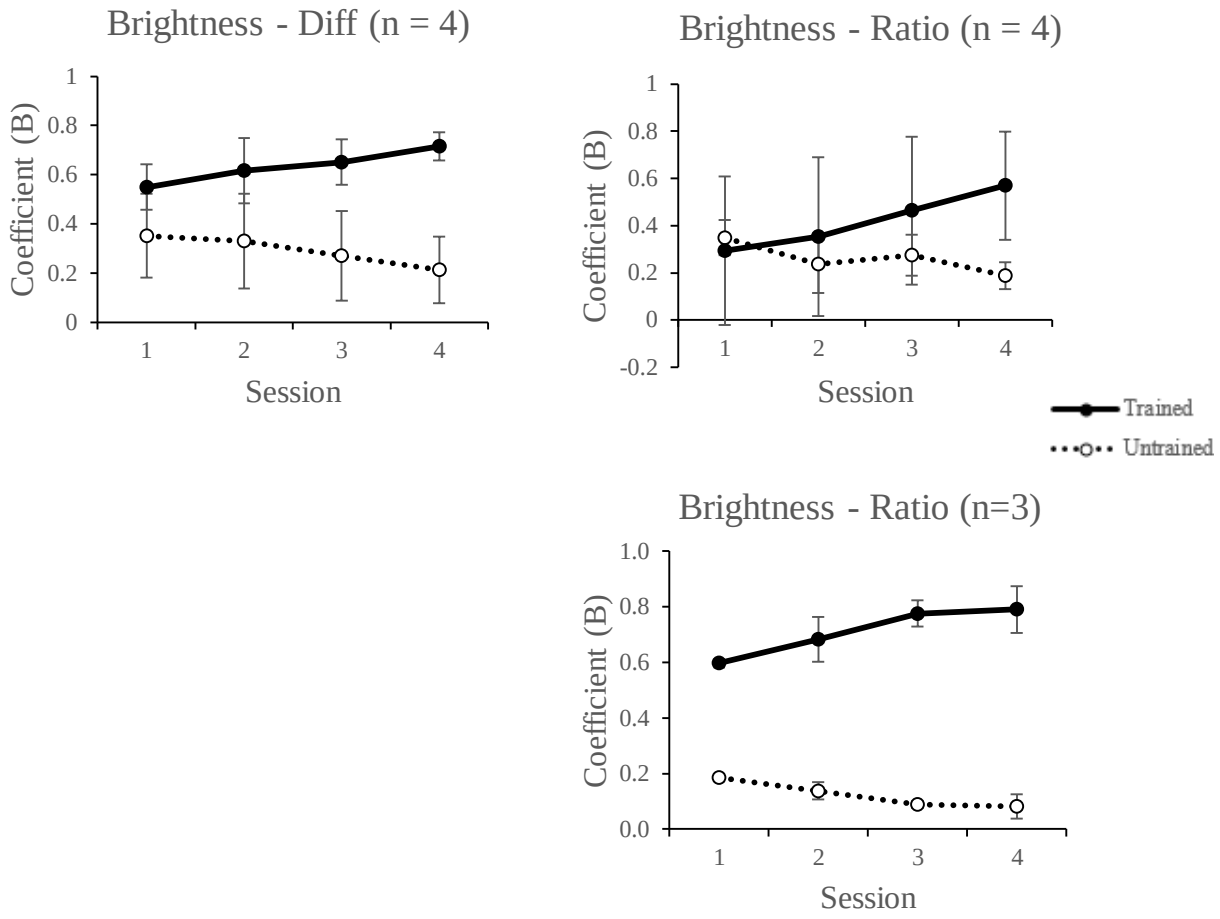

Figure 4. Multiple regression analyses comparing the difference and ratio groups of Experiment 1. The bottom right panel excludes observer Wa.

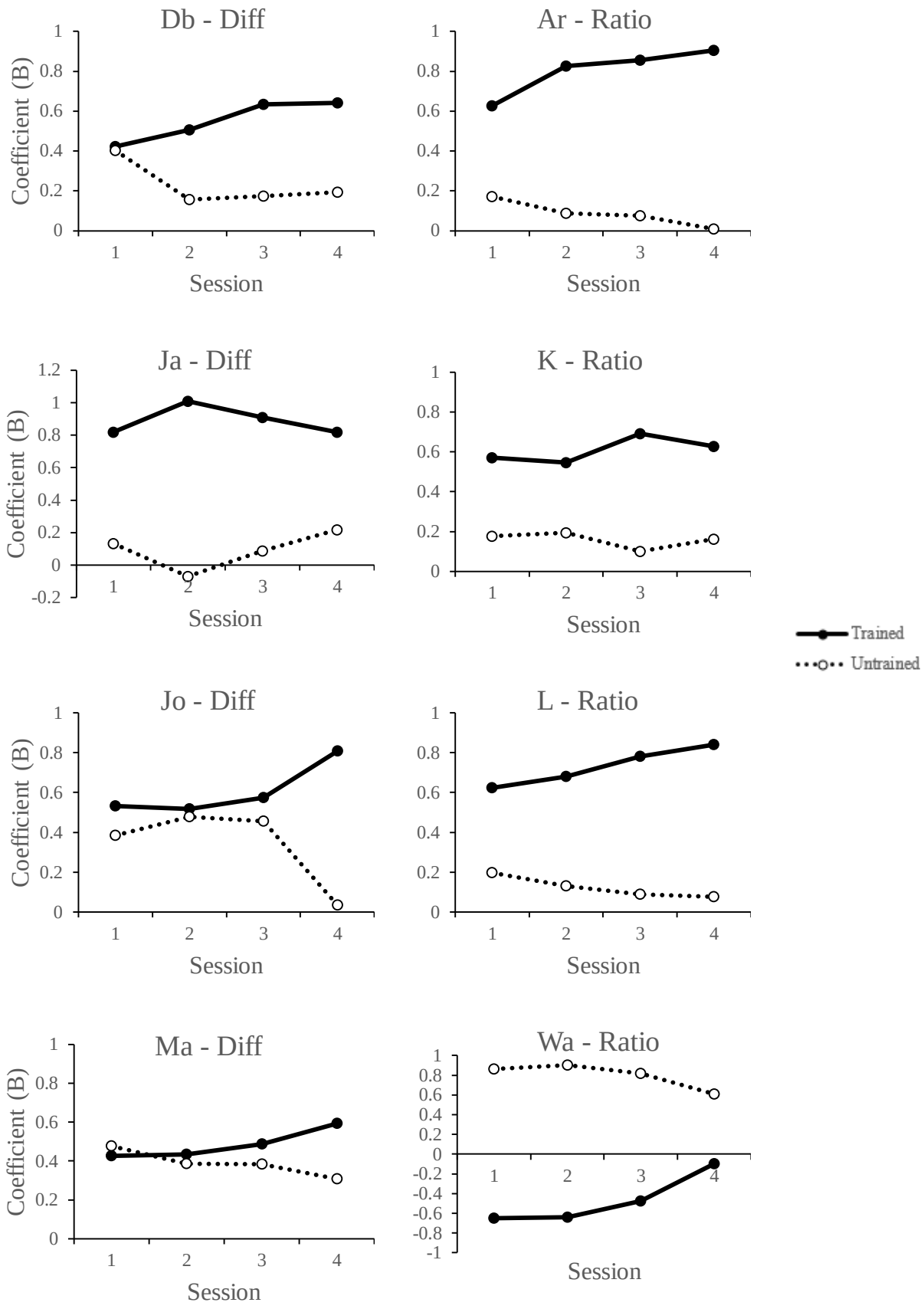

Figure 5. Individual multiple regression analyses for Experiment 1.

A more detailed report of the multiple regression results is shown in Table 2. For each observer, unstandardized regression weights of both relations and overall R^2s are shown for each session. Presence of significant control by the difference or ratio coefficients are noted with asterisks (e.g., * for *p*s < 0.05). Among the 32 sessions combined between observers, 21 (65.62%) of them showed significant control ($p < 0.05$) by both relations. Additionally, all observers had shown significant control by both relations in at least one session. These results suggest that responding cannot be accounted by solely one comparison operation. Therefore, evidence of the present experiment suggest Torgerson's conjecture is false. In line with previous analyses on measures of accuracy, R^2 increased in the fourth session from the first session for all observers. Over sessions, observers' responding generally became more closely aligned with the trained relation.

Table 2. Results of multiple regression for Experiment 1. Complied for both groups is each observer's averaged regression coefficients (unstandardized) for difference and ratio values predicting responding and R^2 for each session.

Difference Group					Ratio Group				
Observer	Session	Diff b	Ratio b	R^2	Observer	Session	Diff b	Ratio b	R^2
Db	1	0.423***	0.402***	0.615	Ar	1	0.171*	0.627***	0.569
	2	0.506***	0.157	0.581		2	0.089	0.826***	0.807
	3	0.635***	0.173*	0.697		3	0.076	0.855***	0.838
	4	0.642***	0.194*	0.669		4	0.008	0.904***	0.855
Ja	1	0.817***	0.132	0.753	K	1	0.177***	0.571***	0.775
	2	1.007***	-0.070	0.844		2	0.195***	0.547***	0.810
	3	0.910***	0.087	0.906		3	0.101*	0.690***	0.806
	4	0.818***	0.218***	0.916		4	0.162**	0.627***	0.802
Jo	1	0.533***	0.385***	0.817	L	1	0.197**	0.623***	0.795
	2	0.519***	0.478***	0.856		2	0.130**	0.679***	0.818
	3	0.575***	0.456***	0.881		3	0.089	0.782***	0.843
	4	0.809***	0.035	0.854		4	0.076	0.841***	0.867
Ma	1	0.425***	0.478***	0.684	Wa	1	0.863***	-0.647***	0.086
	2	0.434***	0.387***	0.638		2	0.905***	-0.641***	0.154
	3	0.486***	0.383***	0.737		3	0.818***	-0.474***	0.287
	4	0.593***	0.308***	0.775		4	0.607***	-0.095	0.491

* = $p < 0.05$
** = $p < 0.01$
*** = $p < 0.001$

Finally, we examined whether patterns of variability in responding were more consistent with predictions of the ANS or OTS. As previously noted, the ANS and OTS differ in response variability. To recapitulate, observers can accurately (and with approximately constant variability) determine numerosity with the OTS automatically, but only for very small numbers. For larger numbers (e.g. n > 5), the ANS is used. The ANS estimates numerosity with less accuracy, and its variability is scalar. For the variability analyses, the trained correct value for each stimulus pair was sorted into one of 20 bins by truncating to intervals of .05. Because the stimulus values were sampled independently from uniform distributions, the distributions of correct values based on differences and ratios were positively skewed, and so the binned intervals did not contain equal numbers of trials. The standard deviations for each bin were calculated for each observer and session, and then averaged across sessions. Results are presented in Figure 6. The two groups resembled each other. A repeated measures ANOVA with session as within subjects factor and group (difference vs ratio condition) as between subjects factor revealed that session was a significant factor, $F(3, 18) = 4.376$, $p = .018$. There was no effect of group or the group x session interaction. Variation of responses decreased over sessions, which means observers responded closer to the mean (I.e. the trained correct value) and observers had improved over sessions. Thus, these results show that responding became more consistent after session 1, but did not change substantially thereafter.

To address the question of whether patterns of variability in the data were more consistent with responding based on the ANS or OTS, the standard deviation for each bin, averaged across subjects for each group, were examined for systematic trends with polynomial regression. Only data from the final session was used. Figure 7 presents these results. The ANS would predict a linear trend among standard deviations, whereas the OTS would suggest a horizontal flat line (i.e., constant variability). However, neither one of those was observed. Instead, an inverted-U shaped curve was found for both difference and ratio groups. The best

fitting trend was selected via hierarchical regression by entering the linear component at the first step, and the quadratic term at the second step, and vice versa. The binned trained values were centred for the linear component and then squared for the quadratic component, to avoid multicollinearity. The linear trend produced a poor fit for the difference group, $R^2 = .039$, $F(1, 17) = .699$, $p = .699$, and the model was significantly improved by the addition of a quadratic component, $R^2 = .715$, $F(2, 16) = 20.038$, $p < .001$. The change between the two models was significant, $F(1, 16) = 37.862$, $p < .001$, with $\Delta R^2 = .675$. When the quadratic term was entered at the first step, the results were significant, $R^2 = .675$, $F(1, 17) = 35.338$, $p < .001$, but the change was non-significant when the linear component was added, $F(1, 16) = 2.214$, $p < .156$, with $\Delta R^2 = .039$. As a result, the quadratic model (without linear component) was taken as the best fitting model for the difference group. For the ratio group, the linear trend produced a poor fit, $R^2 = .130$, $F(1, 15) = 2.242$, $p = .155$. The combined model was significant, $R^2 = .813$, $F(2, 14) = 30.452$, $p < .001$, and so was the change, $F(1, 14) = 51.165$, $p < .001$, with $\Delta R^2 = .683$. The quadratic component by itself was significant, $R^2 = .683$, $F(1, 15) = 32.330$, $p < .001$, and the change after adding in the linear component was also significant, $F(1, 14) = 9.739$, $p = .008$, with $\Delta R^2 = .130$. Therefore, the combined model was taken for the ratio group. Regardless of the chosen model, the inverted-U shaped curve was confirmed by significant quadratic trends in both the difference and ratio group.

One possibility is that stimulus pairs with correct values near both ends of the response bar are easier to compare than correct values closer to the middle. The latter is speculated to be more difficult as these stimulus pairs were somewhat ambiguous in terms of dissimilarity. It is also likely that the bounded nature of the response bar limited response variability near both ends, producing an artefact which resulted in a quadratic trend. Because there was no evidence of an increasing linear trend as predicted by the ANS, the results are most consistent with an assumption of constant variability plus an artefact of response compression due to the lower

and upper response bar limits. Therefore, it is concluded that the present data conforms to the OTS more than the ANS, which is in line with our hypothesis.

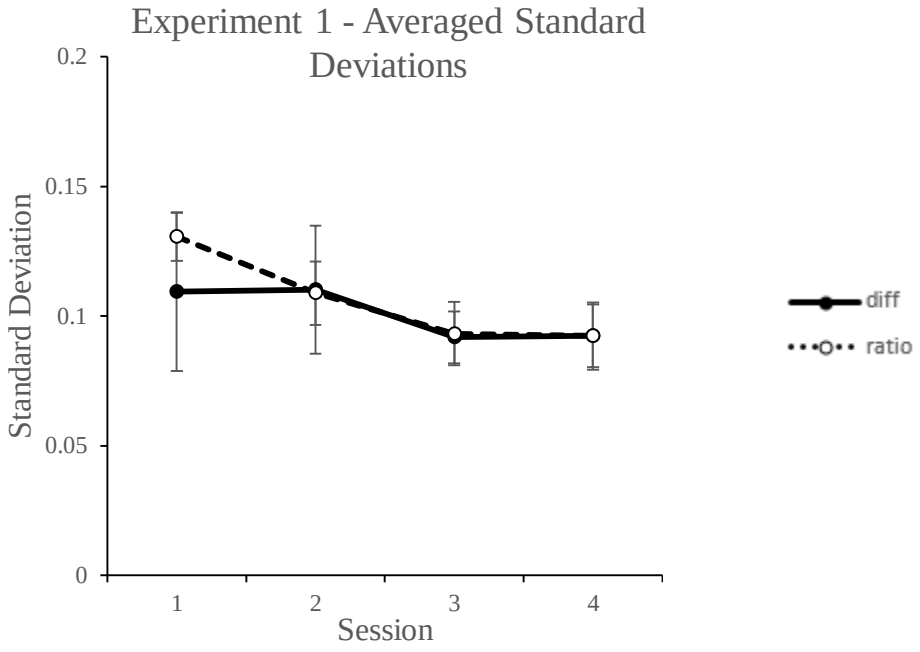

Figure 6. Averaged standard deviations across sessions for Experiment 1. The solid line represents the difference group, whereas the dotted line represents the ratio group.

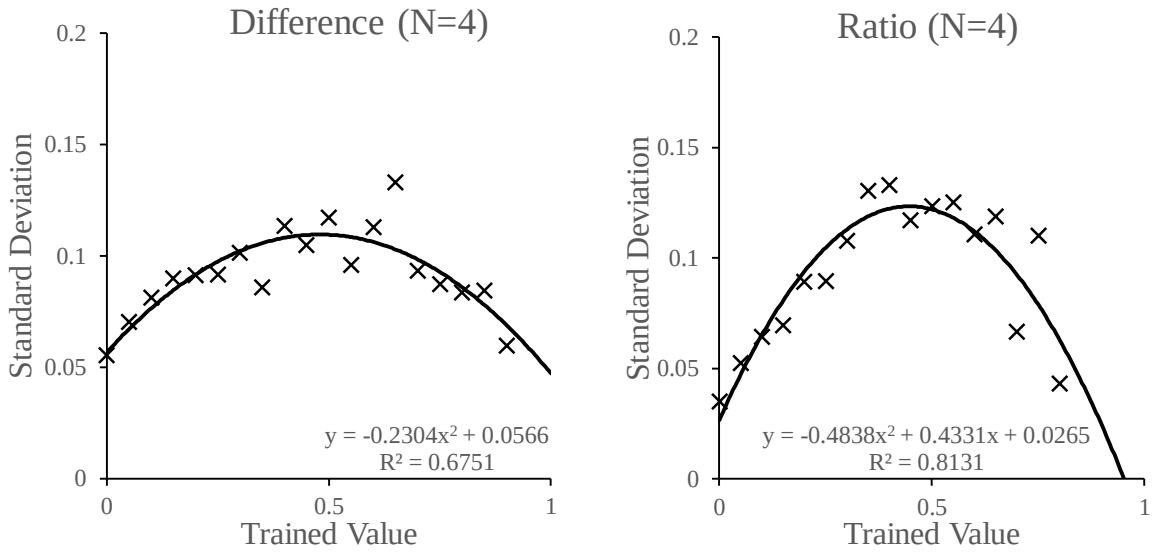

Figure 7. Variability analysis of Experiment 1 (brightness). The right panel shows data of the difference group whereas the left panel shows the ratio group. The binned intervals on the x-axis was rescaled from 0 to 1, with widths of .05. Solid line represents line of best fit, regression equation included. Note: only data from session 4 were included.

Experiment 2

The primary goal of Experiment 2 was to test Grace et al.'s procedure with a new stimulus modality, namely line length. Previous studies have found that subjects are able to accurately judge line lengths in a magnitude estimation task (e.g. Parker, Schneider & Kanow, 1975). This offers a new modality to be tested in Grace et al.'s task. Stimuli modalities have been categorized an extensive if it changes spatially, and as intensive if it changes nonspatially. Recent research has found differing results between extensive and intensive stimuli in a ratio estimation task (Masin, Brancaccio, & Tomassetti, 2019). Hence, it would be interesting to see if the pattern of change across sessions using an extensive dimension (i.e., line lengths) will be similar to that observed with an intensive dimension (i.e., brightness) in Experiment 1. Otherwise, Experiment 2 adopted a similar design and the results are analysed similarly as in Experiment 1.

Method

Participants. Eight observers were recruited for participation (3m, 5f). For Experiment 2, B, Mm, Mi, and R were in the difference condition, and Al, An, De, Wi were in the ratio condition. All were naïve to the purpose of the research, and received either a NZ$60 voucher or course credits for completing all four sessions. All had reported normal or corrected-to-normal vision.

Materials. Pairs of straight lines with differing lengths were used as stimuli. The lines measured 1 mm in width. The lengths varied between 50 and 350 and were always integers, equating to 22 mm to 154 mm in length. The lengths were sampled randomly to make the stimulus sets for each session. The midpoints of the lines were 18 cm apart and 10 cm from the top of the screen. Lines were yellow (R255, G255, B0) to avoid confusion with the response bar. Stimuli were displayed side-by-side and the shorter line was always on the observer's left side. Additionally, the lines were tilted at randomised angles with a difference of at least 30°. This was made to prevent strategies such as directly comparing the length of the lines (such as subtracting the length of the two lines), which would bias responses based on stimuli differences. As both luminance and line lengths can be defined numerically, correct responses are defined as those within the range of ± 7% of the trained value as in the same way as Experiment 1. The response bar, background, and feedback oval in Experiment 2 were all identical to Experiment 1.

Procedure. The procedure for Experiment 2 was the same as Experiment 1. The instructions were similar but referred to line lengths.

Results

As previously discussed, four statistical measures were examined to gauge performance. These measures were investigated in similar manner for Experiment 2. Figure 8 displays the results of the two groups in Experiment 2. The overall trends were similar, as both groups showed responding became more accurate over sessions.

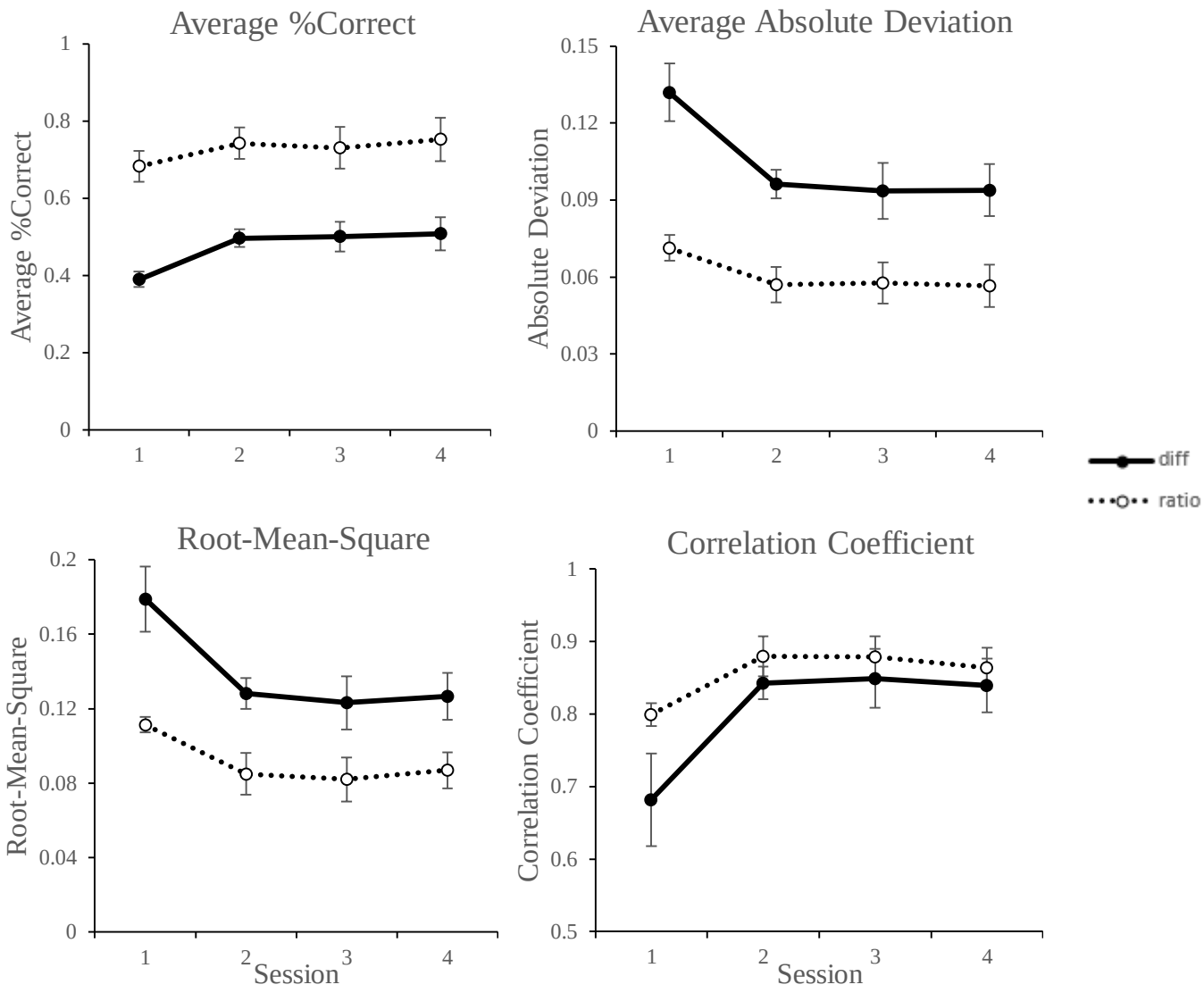

Figure 8. Four measures of accuracy for Experiment 2, average %correct, average absolute deviation, RMS, and R. The solid line represents the difference group (N=4), whereas the dotted line represents the ratio group (N=4).

For average percent correct, the effect of session, $F(3, 18) = 16.006$, $p < .00$, $\eta_p^2 = .727$, and group, $F(1, 6) = 21.086$, $p < .00$, $\eta_p^2 = .778$, were both significant. The interaction between session and group was non-significant ($p = .236$). Post-hoc analyses using the Newman-Keuls test confirmed that the average percent correct trials for Session 1 were significantly lower than for Sessions 2, 3, and 4. From the first to final session, percent correct improved for both groups. The ratio group produced more accurate responses compared to the difference group. This

between-groups significance will be commonly observed in Experiment 2 and will be discussed in more detail later.

For absolute deviations, the effects of session, $F(3, 18) = 24.186$, $p < .00$, $\eta_p^2 = .801$, group, $F(1, 6) = 14.795$, $p < .01$, $\eta_p^2 = .711$, and interaction $F(3, 18) = 4.938$, $p < .01$, $\eta_p^2 = .451$, were all statistically significant. Post-hoc tests (Newman-Keuls) supported that averaged deviations in Session 1 were significantly higher when compared to Sessions 2, 3, and 4. For RMS, both groups showed a reduction in error across sessions. A statistically significant effect was found for RMS in both session, $F(3, 18) = 18.119$, $p < .00$, $\eta_p^2 = .751$ and group, $F(1, 6) = 10.795$, $p < .17$, $\eta_p^2 = .643$. The interaction was not statistically significant ($p = .159$). Post-hoc tests (Newman-Keuls) confirmed that the averaged RMS values for Session 1 were significantly lower than for those in Session 2, 3, and 4. Correlations between responses and correct values significantly increased across sessions, $F(3, 18) = 14.117$, $p < .00$, $\eta_p^2 = .702$. There was no significant effect of group and interaction, ($ps > .164$). Post-hoc tests (Newman-Keuls) confirmed that the averaged correlations in Session 1 were significantly less than the averages for Sessions 2, 3, and 4.

Using the results of Experiment 2, Table 3 shows the averaged means and standard deviations of sessions 1 and 4 of the four measures. Same as the previous experiment, both groups had shown an improvement for all four measures. These results show that extended training sessions had served a beneficial effect on performance. It is noted that these improvements were stronger for the difference group in Experiment 2, whereas they were stronger for the ratio group in Experiment 1.

Table 3. Results of four measures of accuracy of Experiment 2 (line lengths) difference and ratio group. Listed for each measure are the averaged means and standard deviations for the trained relation in sessions 1 and 4. The % improved represents the percentage increase for percentage correct and r, and percentage decrease for absolute deviation and RMS.

Difference Group					Ratio Group				
Measure	Session	Mean	SD	% improved	Measure	Session	Mean	SD	% improved
%Corr	1	0.390	0.040	30.153	%Corr	1	0.68229	0.080	10.251
	4	0.507	0.086			4	0.752	0.112	
AbsDev	1	0.132	0.023	28.891	AbsDev	1	0.071	0.010	20.585
	4	0.094	0.020			4	0.057	0.017	
RMS	1	0.179	0.035	29.166	RMS	1	0.111	0.008	22.019
	4	0.127	0.025			4	0.087	0.020	
r	1	0.681	0.128	23.195	r	1	0.799	0.031	8.090
	4	0.840	0.074			4	0.863	0.057	

Figure 8 is a replica of Figure 1 using the results of Experiment 2. They both point toward to same general trends, which further consolidates our previous conclusions of the beneficial effect of extended training. In Experiment 2, the results were more robust. The error bars were wider for Experiment 1, and this was due to one anomalous observer in the ratio group (Wa). This particular observer responded based on differences initially, then slowly used more of the ratios between stimuli over sessions. Hence, the weaker between-group differences in the brightness experiment are due to observer Wa as well. However, the overall trends are still similar in both experiments – measures of accuracy increased over sessions.

Figures 9 & 10 shows representative scatterplots of the trained values plotted against response. They each represent the difference and ratio group for Experiment 2, respectively. Figures 9 & 10 were generated in the same way as Figures 2 & 3. Responses were scattered around the line of perfect positive correlation (as represented by the solid line). Consistent across the two experiments, the scaled trained values were strong predictors of responses in both conditions. This was observed as early as in the first sessions. Even without explicit feedback, there was rapid learning of the trained relation with stimuli that varied in brightness or line length. Moreover, correlations of the trained relation strengthened over sessions. Out of 16 participants, 15 (except Db; Figure 10, centre column) showed greater R^2 in the final session compared to the first session. This suggests repetitious training of the task was able to improve performance. This is consistent with our previous findings of observers showing improvement over sessions.

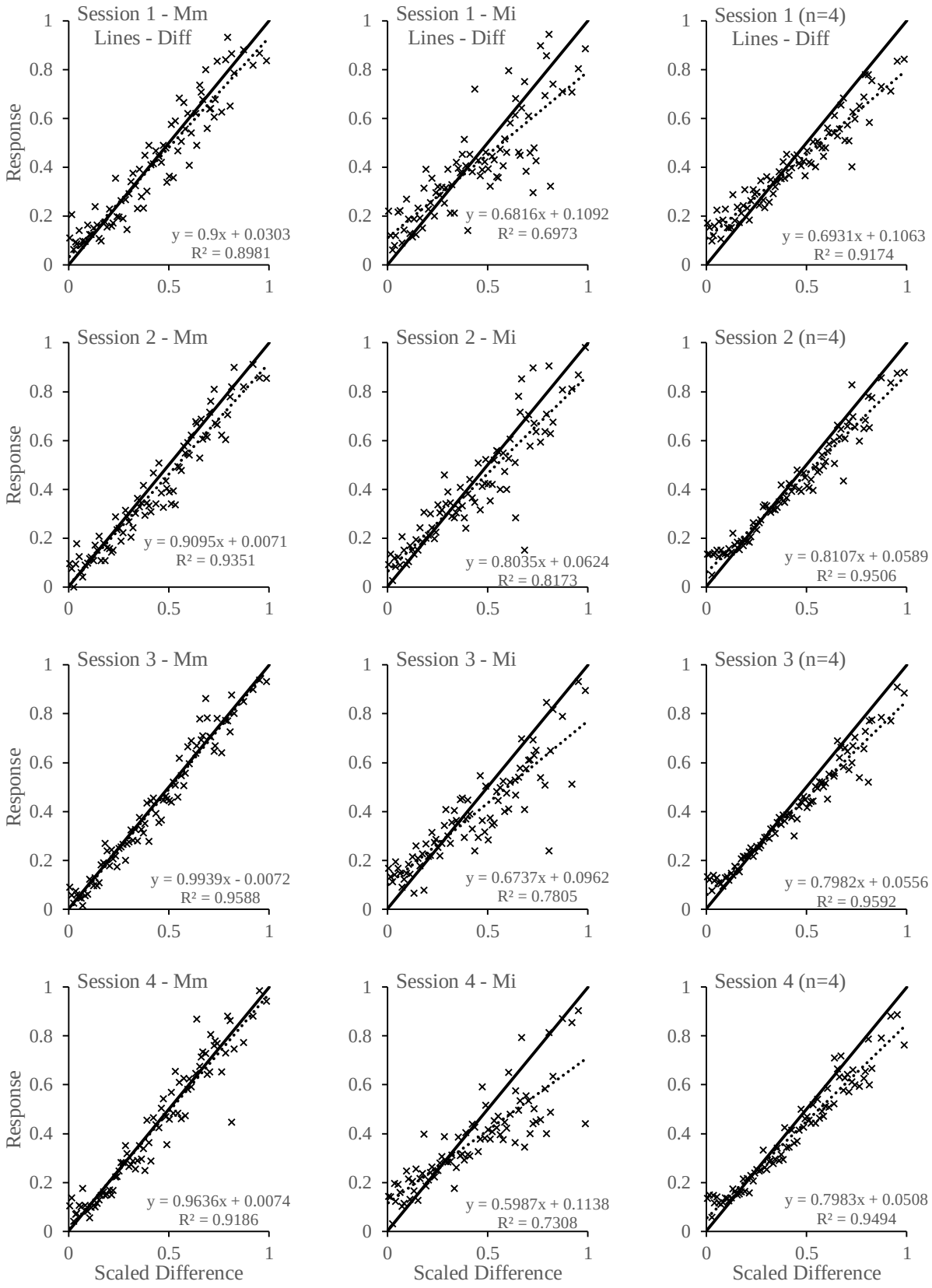

Figure 9. Scatterplots of Experiment 2 difference group as a measure of accuracy. The solid line indicates perfect correlation; the dotted line indicates line of best fit. From left to right shows the best and worst performer, and the averaged results, respectively.

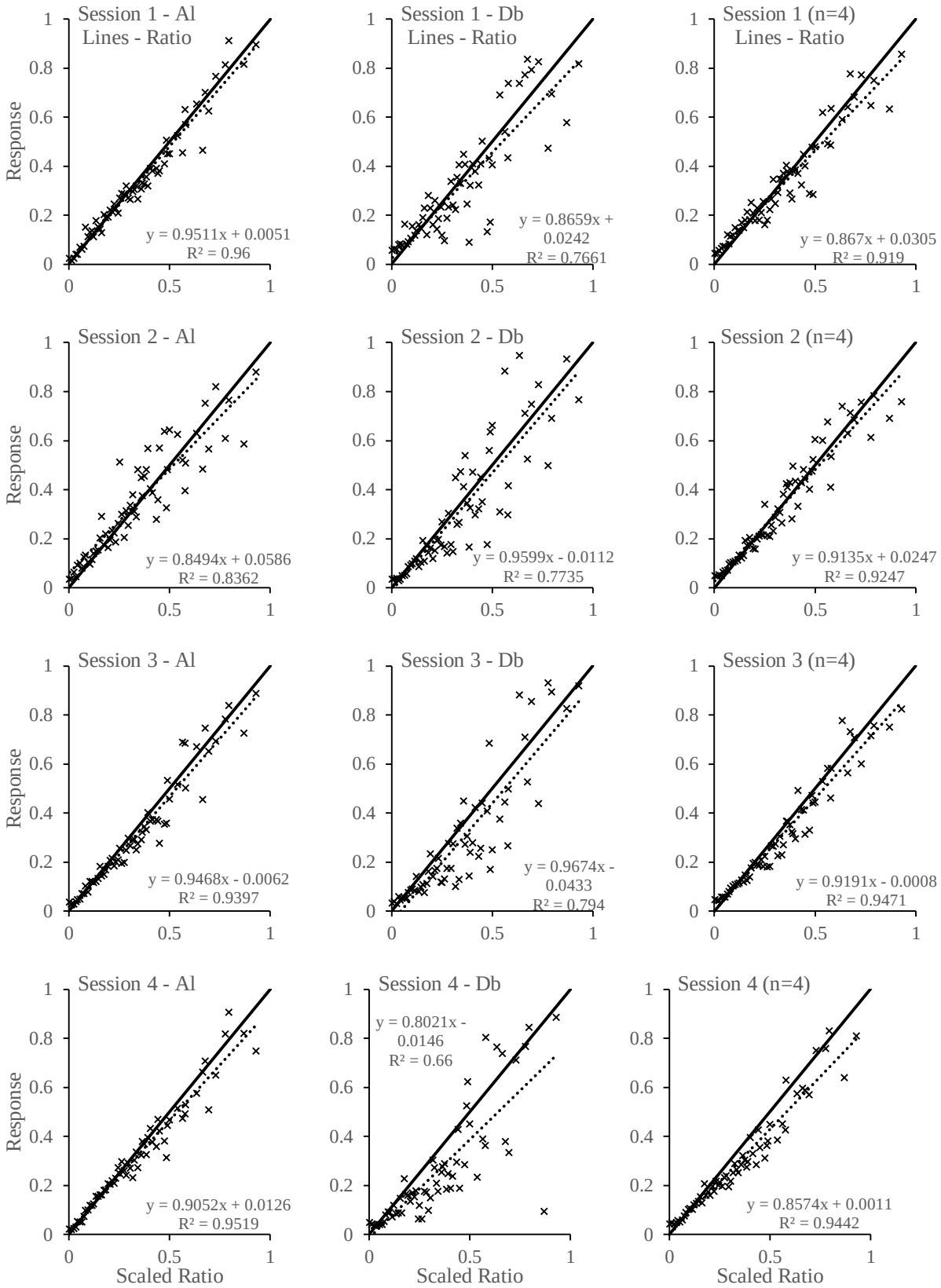

Figure 10. Scatterplots of Experiment 2 ratio group as a measure of accuracy. The solid line indicates perfect correlation; the dotted line indicates line of best fit. From left to right shows the best and worst performer, and the averaged results, respectively.

Figure 11 presents the multiple regression trends of Experiment 2, averaged across participants. Both groups displayed greater coefficients for their respective trained relation. Across sessions, participants showed steadily increasing learning of the trained response. The B weights for the untrained relation decreased over sessions. A repeated measures ANOVA with session and relation (trained vs untrained) as within subjects factors, and group (difference vs ratio) as between subjects factor was examined. Same as Experiment 1, session, $F(3, 18) = 5.225$, $p = .009$, and the session x relation interaction, $F(3, 18) = 3.555$, $p = .035$, was found to be significant. Additionally, relation yielded a statistically significant result, $F(1, 6) = 68.779$, $p < .000$. Session had the opposite effect on the trained and untrained relation – the trained relation was utilized more and more over sessions, whereas the untrained relation showed a decreasing trend over sessions. These results are in line with the findings of Experiment 1. This further reinforces our finding of observers' ability to learn the trained relation when given simple feedback without explicit instructions. The averaged results are representative of individual data. Figure 12 shows the individual multiple regression trends of Experiment 2.

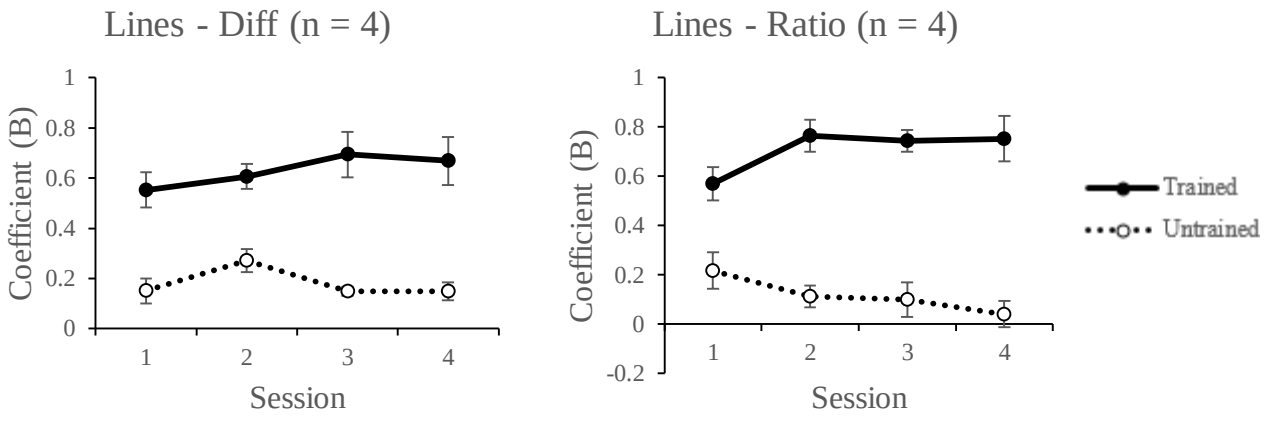

Figure 11. Multiple regression analyses of the difference and ratio conditions of Experiment 2.

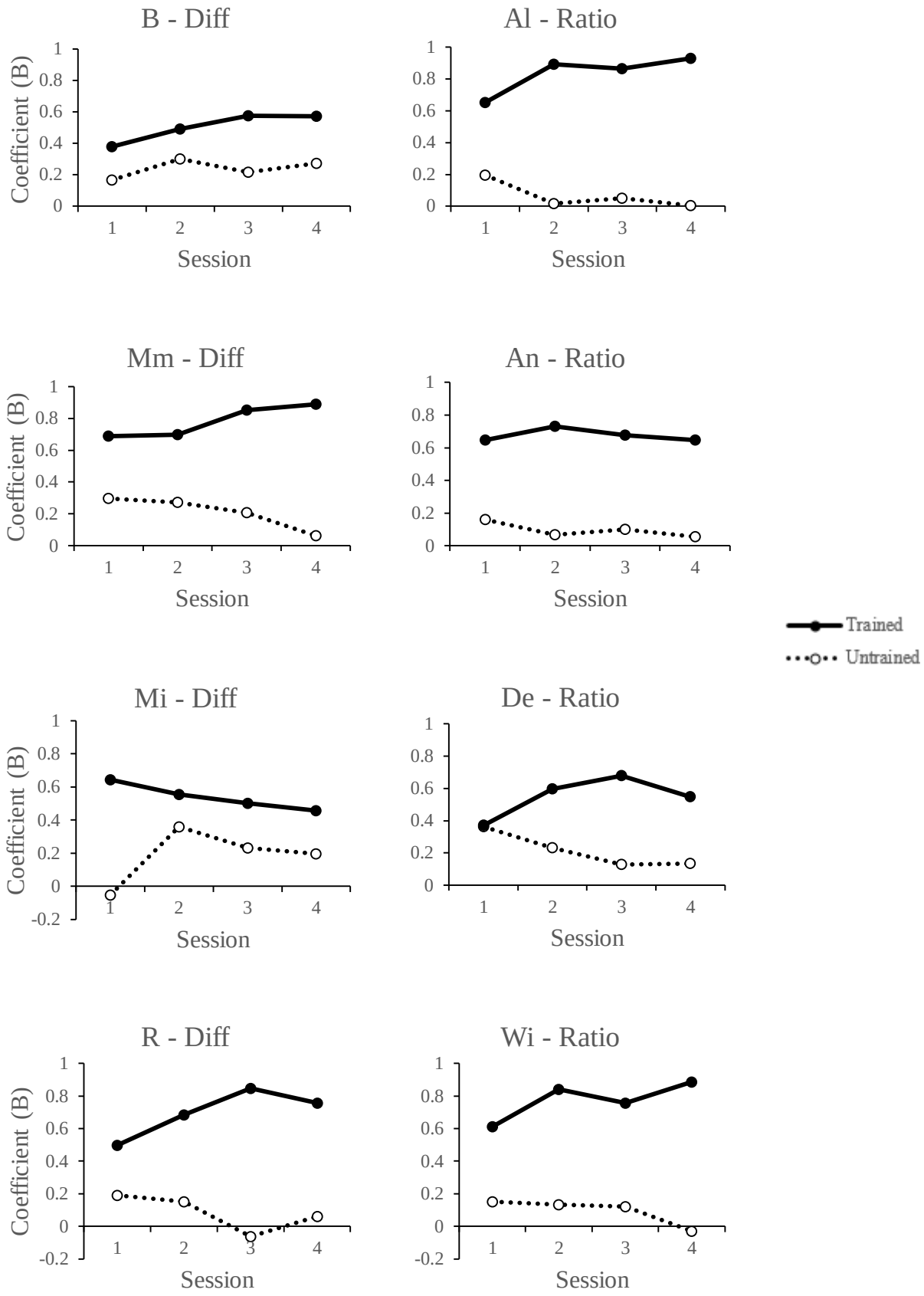

Figure 12. Individual multiple regression analyses for Experiment 2.

Detailed results of the multiple regression analyses are shown in Table 4. Of the 32 sessions in Experiment 2, 21 (65.62%) of them displayed significant control ($p < 0.05$) by both differences and ratios. Furthermore, observers were able to respond well to the trained feedback. All 32 sessions had shown strong ($p < 0.001$) control of the trained relation. R^2 also increased over sessions for all observers. Observers were able to respond closer towards the trained relation after given training. Combining both experiments, all but one observer (Wa; Figure 5, bottom right panel) showed greater coefficients for the trained relation compared to the untrained relation in all four sessions. In addition, all observers displayed significant control ($p < 0.05$) by the trained relation in all but one session (Wa in the 4th session). Despite one outlier participant, observers were successful in adopting to the feedback and respond accordingly.

Table 4. Results of multiple regression for Experiment 2. Complied for both groups is each observer's averaged regression coefficients (unstandardized) for difference and ratio values predicting responding and R^2 for each session.

Difference Group					Ratio Group				
Observer	Session	Diff b	Ratio b	R^2	Observer	Session	Diff b	Ratio b	R^2
B	1	0.380***	0.167	0.291	Al	1	0.196***	0.652***	0.653
	2	0.492***	0.302***	0.647		2	0.017	0.892***	0.898
	3	0.577***	0.216*	0.723		3	0.049*	0.863***	0.922
	4	0.573***	0.274***	0.725		4	0.003	0.929***	0.870
Mm	1	0.689***	0.295***	0.713	An	1	0.161***	0.646***	0.704
	2	0.696***	0.271***	0.782		2	0.067	0.731***	0.746
	3	0.850***	0.207**	0.848		3	0.098*	0.678***	0.724
	4	0.889***	0.062	0.834		4	0.056	0.647***	0.718
Mi	1	0.644***	-0.052	0.409	De	1	0.364***	0.370***	0.599
	2	0.554***	0.361***	0.647		2	0.230***	0.596***	0.665
	3	0.500***	0.233*	0.541		3	0.128**	0.678***	0.698
	4	0.458***	0.195*	0.542		4	0.134**	0.548***	0.635
R	1	0.498***	0.190	0.493	Wi	1	0.151**	0.613***	0.599
	2	0.683***	0.152*	0.771		2	0.134**	0.842***	0.795
	3	0.848***	-0.061	0.791		3	0.122**	0.757***	0.755
	4	0.755***	0.060	0.735		4	-0.029	0.887***	0.768

* = $p < 0.05$
** = $p < 0.01$
*** = $p < 0.001$

Patterns of variability in responding were analysed for Experiment 2 in the same way as Experiment 1. Figure 13 presents the averaged standard deviations over sessions. A repeated measures ANOVA with session as within subjects factor and group (difference vs ratio condition) as between subjects factor found a significant effect of session, $F(3, 18) = 9.955$, $p < .00$. The effects of group and session x group interaction were not statistically significant. Consistent with Experiment 1, response variability decreased over sessions.

For each group, Figure 14 shows the averaged standard deviations for the correct value bins. The inverted-U shaped curvature observed in Experiment 1 was less obvious here. Hierarchical regression was conducted to find the best fitting trend for each group. A linear component was entered at the first step and a quadratic component was added in at the second step, as per Experiment 1. The difference group was best characterized by a linear trend, $R^2 = .197$, $F(1, 18) = 4.403$, $p = .050$, as entering the quadratic term produced non-significant results, $R^2 = .251$, $F(2, 17) = 2.845$, $p = .086$, and the change was also non-significant, $F(1, 17) = 1.231$, $p < .283$, with $\Delta R^2 = .054$. Non-significant results were obtained when the quadratic term was entered at the first step, $R^2 = .021$, $F(1, 18) = .381$, $p = .545$. Hence, the linear model was taken as the best fitting model, despite the relatively low R^2 (.197) that approached significance ($p = .050$). For the ratio group, the linear model was significant, $R^2 = .382$, $F(1, 12) = 7.412$, $p = .019$, whereas the combined model also showed significant results, $R^2 = .433$, $F(2, 11) = 4.209$, $p = .044$. However, the change was non-significant, $F(1, 11) = 1.004$, $p = .338$, with $\Delta R^2 = .052$. The results were non-significant when the quadratic component was first added, $R^2 = .007$, $F(1, 12) = 0.079$, $p = .783$, and the change when adding the linear component was significant, $F(1, 11) = 8.29$, $p = .015$, with $\Delta R^2 = .433$. Therefore, the linear model was chosen for the ratio group. Although Experiment 2 showed mild support in favour of the ANS in our perceptual comparison task, the effect sizes were lower in support of the ANS ($R^2s = .197$ and .382 for the difference and ratio group, respectively), compared to results from Experiment

1. Moreover, outliers were present in both groups of Experiment 2, which may have produced bias for the linear model. Excluding the most significant outlier for each group (marked bold and colored red in Figure 14), both the difference and ratio group favored the combined model, for the difference group, $R^2 = .515$, $F(2, 16) = 8.480$, $p = .003$, and for the ratio group, $R^2 = .498$, $F(2, 10) = 4.969$, $p = .032$. After omitting the outliers, the changes from a linear to a combined model became statistically significant in both groups.

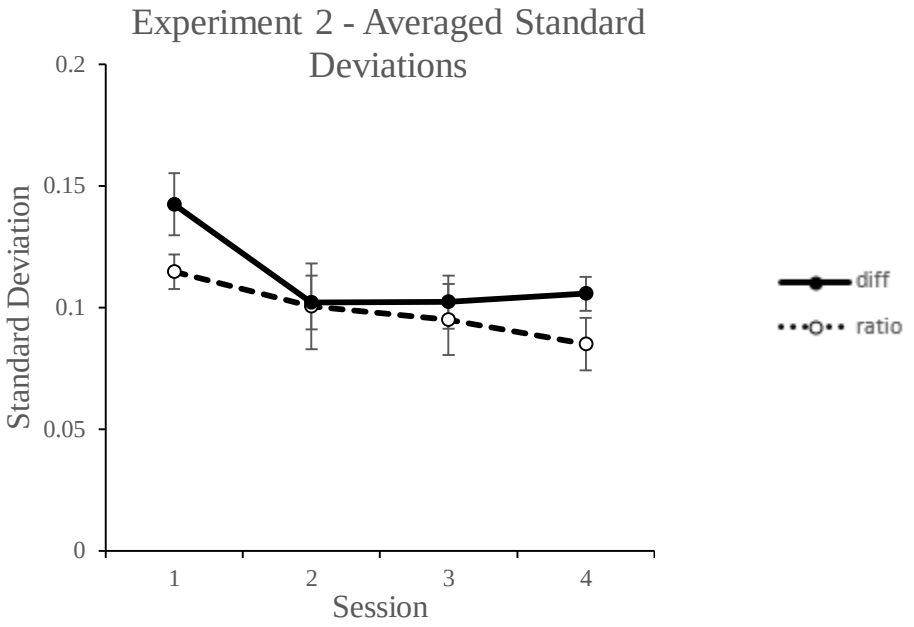

Figure 13. Averaged standard deviations across sessions for Experiment 2. The solid line represents the difference group, whereas the dotted line represents the ratio group.

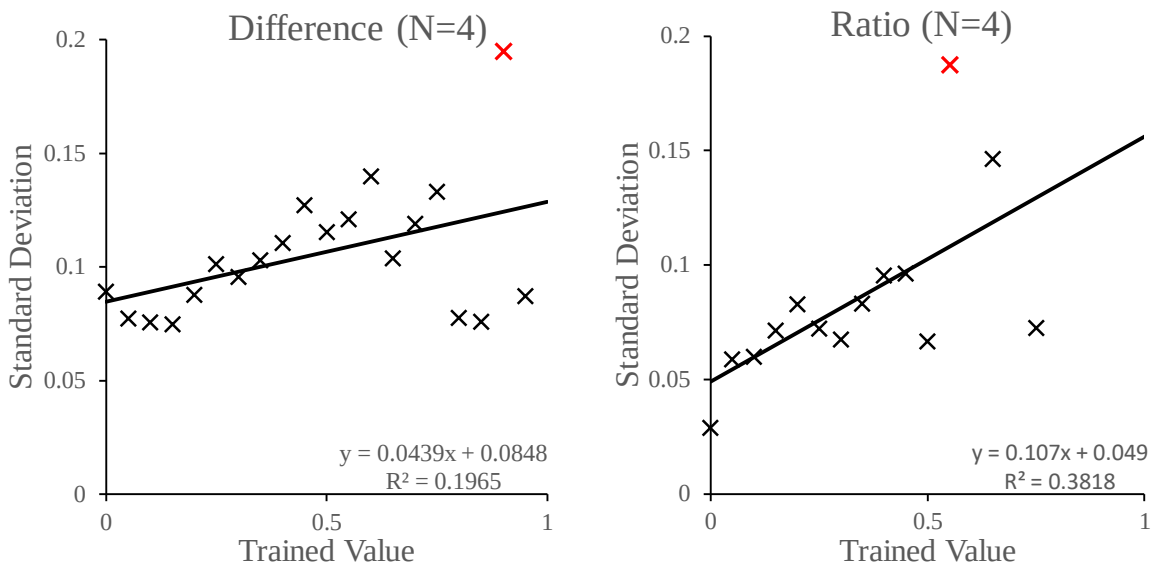

Figure 14. Variability analysis of Experiment 2 (line lengths). The right panel shows data of the difference group whereas the left panel shows the ratio group. The binned intervals on the x-axis was rescaled from 0 to 1, with widths of .05. Solid line represents line of best fit, regression equation included. Note: only data from session 4 were included. Most significant outlier for each group is bolded and colored red.

General Discussion

The overall aim of the present research was to investigate how human observers compare perceptual magnitudes. According to Torgerson's conjecture (Torgerson, 1961), observers may only perceive a single quantitative relation between stimuli, and either subjective differences or subjective ratios is used to compare stimulus magnitudes. Subsequent studies have tested Torgerson's conjecture for a wide array of modalities, and provided insight on how observers respond when asked to estimate differences and ratios numerically (Birnbaum, 1982). More recently, Grace et al. (2018) developed a magnitude comparison task which did not include numbers or explicit instructions on the basis that it would be more fundamental to investigate Torgerson's conjecture when numbers and explicit cognitive processes are not involved. This led to the two main differences which set Grace et al.'s (2018) procedure apart: 1. instead of providing instructions on how to perform the task, observers were trained via feedback, and 2. instead of numerical responding, observers responded by clicking on a continuous bar below the stimuli. Their results showed observers were able to learn the trained arithmetic relations. Averaged across experiments, response and trained values were strongly correlated, *r*s = .95 and .94, for the difference and ratio condition respectively. More importantly, significant control over responding was found for both differences and ratios, regardless of which was trained. Based on this finding, it was concluded that Torgerson's conjecture is false.

The present study serves as a follow-up to Grace et al. (2018) with aims to see if their results could be replicated with a new modality (line length) and with additional training (four sessions), and to revisit the predictions of Torgerson's conjecture. A methodology which was

based around the design of Grace et al. (2018) was adopted, by again using the implicit, nonsymbolic task. Modalities which were used by Grace et al. (brightness; in Experiments 1a & 2a) and newly introduced (line lengths) were tested in Experiment 1 and Experiment 2, respectively. Although the problem of exemplar learning was addressed in Grace et al. (2018), we countered this by randomly generating stimulus values, rather than using a limited number of pairs that were repeatedly presented, as in Grace et al.'s study. Furthermore, the effect of training was examined by administering extended sessions – each observer attended a total of four sessions. Based on contemporary literature on perceptual and reinforcement learning, it was hypothesized that observers' responding would become more closely aligned with the trained relation over sessions – that is, over sessions, control by the trained relation would increase and control by the untrained relation would decrease.

Observers in both experiments showed successful learning of their assigned trained relation. Averaged across experiments and the four sessions, correlations between response and trained value were strong, *r*s = .84 and .82, respectively. Therefore, observers captured learning of the trained relation and responded accurately. In addition, accuracy measures indicated that responding became more accurate across sessions (Figures 1 & 8). Comparing session 1 to session 4, all four measures of accuracy suggested that observers had improved in performance over the course of training (Tables 1 & 3). These improvements were significant for both groups in both experiments. However, most of these improvements occurred in the first to second session, with little improvement thereafter. Averaged data were representative of individuals. Individual scatterplots supported these conclusions, as the positive correlations between response and scaled trained relations were strong, and increased over sessions (Figures 2, 3, 9, 10). These results are consistent with those of Grace et al. (2018).

In Experiment 1, there were no significant differences in accuracy depending on whether observers were trained with difference or ratio relations, however, repeated measures

ANOVAs revealed between groups difference in Experiment 2, where the ratio group outperformed the difference group in three out of the four measures of accuracy. Whilst it is speculative, it may be that line lengths are more easily judged with ratios rather than differences. However, individual differences are also a likely candidate for this finding due to the relatively low number of observers in each group (n = 4). Therefore, reasons why we obtained a group difference in accuracy with line length but not with brightness as the stimulus modality are unclear.

Averaged across groups, multiple regression analyses showed control by the trained relation was significantly higher than the untrained relation (Figures 4 & 11). Most importantly, there was an interaction effect in both experiments suggesting that training provided the opposite effect on the trained and untrained relation – whereby observers responded more by the trained relation over sessions, control by the untrained relation decreased. This result suggested that through practice observers were able to learn and respond based on the trained arithmetic relation between stimulus values. Multiple regression analyses of individuals support group data (Figures 5 & 12; Tables 2 & 4).

One key aim was to revisit Torgerson's conjecture – whether perceptual comparisons in an implicit, non-symbolic paradigm would show evidence of one or two operations. Exploring the validity of Torgerson's conjecture, Tables 2 and 4 collated the unstandardized coefficients of Experiments 1 and 2, respectively. Combining both experiments, all observers had shown significant control by both differences and ratios in at least one session. Furthermore, 13 (81%) observers showed significant control by both operations in at least two sessions, of which 7 (43%) had shown joint control in all four sessions. These results do not follow the predictions of Torgerson's conjecture, as observers' responding showed significant control by both differences and ratios when comparing stimuli. This result was found for both brightness and line lengths, and the latter suggested that the conclusion of Grace et al. (2018) could be

extended to an extensive dimension. Under the current circumstances, the present study rejects Torgerson's conjecture; this is consistent with Grace et al.'s (2018) results.

As previously noted, observer Wa produced a pattern of responding that was unusual compared to the others. In the first session, Wa showed a correlation of r = .224 between response and trained value which is drastically lower than the second worst performer of r = .648. In the session Wa also showed complete opposite control by the two relations, as manifested by negative B coefficients (-0.647, $p < 0.001$) in the trained relation, and significant control (B = 0.863, $p < 0.001$) by the untrained relation. Observer Wa's responding was more strongly controlled by differences than ratios, even though feedback was based on ratios.

A recent study by Masin, Brancaccio, and Tomassetti (2019) may provide some insight into Observer Wa's results. Masin, Brancaccio, and Tomassetti (2019) tested observers' ability to make ratio estimates of stimuli with extensive and intensive modalities. They defined sensations as extensive, if they change spatially, and as intensive, if they change nonspatially. In their experiments, they found subjects were unable to estimate ratios of brightnesses accurately, but were able to accurately estimate ratios of line lengths. They argued that observers could estimate ratios of pairs of lengths by quickly counting how many times the shorter length is contained in the longer length (Hartley, 1977; Masin, 2013). Studies supported this ability by showing high correlation between ratio judgments of perceived distance and the corresponding ratios of physical distance (Norman, Adkins, & Pedersen, 2016; Masin, Brancaccio, & Tomassetti, 2019). This ratio judging technique is not applicable to brightness estimations. Thus, it was suggested that people can judge ratios of extensive sensory magnitude, but not ratios of intensive sensory magnitude. Observer Wa's apparent difficulty with learning to respond based on ratios of brightness, reflected in the negative B coefficient for ratios, may indicate suppression or inhibition of ratios in an attempt to respond based on differences (Figure 5, bottom right panel). Furthermore, studies have suggested that there persist major

idiosyncratic differences in the speed and accuracy of estimation (Tosto et al., 2017; Halberda, Mazzocco, & Feigenson, 2008). By using different nonsymbolic tasks, individual differences in nonsymbolic estimation have been found in preschool children, school-aged children, and adults (Barth et al., 2006; Gilmore, McCarthy, & Spelke, 2010; Halberda et al., 2012; Nys & Content, 2012). Nonetheless, observer Wa showed steady improvement over sessions, improving to a correlation of r = .595 in the final session compared to r = .660 by the second worst performer in the fourth session. In both correlation and multiple regression analyses (Figures 3 & 5), Wa's response patterns followed the group trend and showed improvement over sessions by increased control of the trained operation. Despite early failure, these findings do suggest signs of learning and improved performance over time.

According to current theory in the field of numerical cognition, our ability to estimate magnitudes include two systems, the ANS and the OTS (Feigenson, Dehaene, & Spelke, 2004). The ANS is for the representation of large, approximate numerical magnitudes, and the OTS is for the precise representation of small numbers of individual objects. We sought whether if data gathered from our task would conform more to one or the other system. As applied to our task, this would manifest in the form of scalar variability responding for the ANS, or constant variability in responding for the OTS. Experiment 1 (Figure 7) displayed inverted-U shapes for the variability, which was confirmed by a significant quadratic trend, whereas Experiment 2 (Figure 14) showed slight linear trends. When the outliers were removed for both groups in Experiment 2, the combined trend (linear and quadratic) became the most fitting instead. Therefore, variability patterns were more inclined to the explanations of the OTS, which suggests an 'automatic' processing of the perceptual comparisons (see also Kahneman, 2011). However, these results were somewhat mixed and should be followed up in future research.

The potential for multicollinearity problems surface when predictors in multiple regression are highly correlated. A general rule of thumb has been to regard standardized

coefficients (beta weights) that are > 1 as potential indicators of multicollinearity problems (Kraha, Turner, Nimon, Zientek, & Henson, 2012). This was found in observer Wa's data in sessions 2 and 3, with beta coefficients for the differences being 1.128 and 1.217, respectively, while coefficients for ratios were negative. Grace et al. (2018) addressed the possibility that results might have been influenced by multicollinearity by using Monte Carlo analyses. Similarly, we conducted Monte Carlo simulations to test retrieval of difference and ratio coefficients using defined weights and noise. Within the range of most observers' B coefficients, the defined weights for the difference and ratio coefficients ranged from 0 to .8, and the noise (σ) ranged from 0.01 to 0.2. However, because observer Wa showed statistically significant negative unstandardized coefficients in several sessions (ratios in sessions 1, 2, and 3; see Tables 2), Monte Carlo simulations of defined weights with negative values were also investigated. The positive coefficients were successfully recovered, suggesting that Tables 2 and 4 provided valid information regarding the relative control by differences and ratios over responding. To address Wa's abnormal data, we conducted additional Monte Carlo simulations of defined weights approximate to Wa's unstandardized coefficients. Results showed that Wa's coefficients were successfully recovered by multiple regression. From a set of 1000 iterations with defined parameters of σ = .05, difference coefficient = .8, and ratio coefficient = -.6, the averaged multiple regressions showed difference coefficient = .782 (SD = .101), ratio coefficient = -.589 (SD = .100), R^2 = .759 (SD = .071), and intercept = .005 (SD = .019). These results suggest that Wa's anomalous result is not due to an artefact of multicollinearity. This supports our hypothesis that the negative coefficient for ratios might suggest that Wa was inhibiting ratios and attempting to respond based on differences, even though feedback was based on ratios.

There are some potential limitations in this study, which should be noted. Firstly, the relatively restricted sample size in each group (n=4) may limit the generality of the results.

This poses possibility for low statistical power, false discovery, and inflated effect size estimation (Button et al., 2013). These issues would undermine the obtained empirical results as low statistical power reduces the chance of detecting a true effect. This also allows for outliers to have a much greater impact on group trends. For example, multiple regression analyses of groups in Experiment 1 showed significantly differing results when Wa's data was omitted (Figure 4). However, doing so sacrificed power as the sample size was low. Furthermore, Grace et al.'s (2018) procedure to investigate perceptual comparisons using implicit, nonsymbolic stimuli is novel, with no prior research. Studies with aims to replicate these results are advised for future research to overcome these problems.

Alternative designs to assess implicit learning processes with nonsymbolic stimuli are also recommended for future research. For example, auditory stimuli have not been tested under the present methodology. Moreover, our observers have mostly been college adults, and longitudinal studies have shown correlations between math achievement and educational attainment (Cortes, Goodman, & Nomi, 2015). The sampled observers presented in this study may not be representative of the general populace, and demographic factors such as age and education level might affect performance on our task. Presenting this task to a wider variety of observers in terms of age and educational background should be followed up in future research. Finally, children before the development of algebra could show interesting implications for our task in a more fundamental sense, as prior to learning maths they have shown capacity to judge numerically differing magnitudes (Barth et al., 2006; Barth, Beckmann, & Spelke, 2008).

Building on the paradigm introduced by Grace et al. (2018), the present study adopted a perceptual comparison procedure that used implicit feedback and nonsymbolic analogues response. There were some alterations, including extended sessions, random stimulus values as opposed to set pairs, and a new modality (line lengths) was introduced. Analyses of results showed quick learning of the task and improvement over sessions in both experiments. Taken

together, the conclusions of this study were in agreement with Grace et al. (2018). Results indicated that observers would apply both differences and ratios between stimuli when comparing stimuli even though observers were trained on a single relation. Therefore, we conclude that Torgerson's conjecture is false, and that the perceptual system computes two operations, corresponding to ratios and differences, when comparing magnitudes.

References

Alvarez, G. A., Cavanagh, P. (2004). The Capacity of Visual Short-Term Memory Is Set Both by Visual Information Load and by Number of Objects. *Psychological Science, 15(2),* 106-111.

Baird, J. C., Green, D. M., & Luce, R. D. (1980). Variability and sequential effects in cross-modality matching of area and loudness. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(2), 277–289. https://doi.org/10.1037/0096-1523.6.2.277

Barth, H., Beckmann, L., & Spelke, E. S. (2008). Nonsymbolic, approximate arithmetic in children: Abstract addition prior to instruction. *Developmental Psychology*, *44*(5), 1466–1477. https://doi.org/10.1037/a0013046

Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., & Spelke, E. (2006). Non-symbolic arithmetic in adults and young children. *Cognition*, *98*(3), 199–222. https://doi.org/10.1016/j.cognition.2004.09.011

Berch, D. B. (2005). Making sense of number sense: implications for children with mathematical disabilities. *Journal of Learning Disabilities*, *38*(4), 333–339. https://doi.org/10.1177/00222194050380040901

Birnbaum, M. H. (1978). Differences and ratios in psychological measurement. *Cognitive Theory, 3*, 33-74.

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 401–485), Hillsdale, NJ: Erlbaum.

Birnbaum, M. H., Anderson, C. J., & Hynan, L. G. (1989). Two operations for "ratios" and "differences" of distances on the mental map. *Journal of Experimental Psychology. Human Perception and Performance*, *15*(4), 785–796.

Birnbaum, M. H., & Veit, C. T. (1974). Scale convergence as a criterion for rescaling: Information integration with difference, ratio, and averaging tasks. *Perception & Psychophysics*, *15*(1), 7–15. https://doi.org/10.3758/BF03205820

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, *42*(1), 189–201. https://doi.org/10.1037/0012-1649.41.6.189

Brennan, E. M., Ryan, E. B., & Dawson, W. E. (1975). Scaling of apparent accentedness by magnitude estimation and sensory modality matching. *Journal of Psycholinguistic Research*, *4*(1), 27–36. https://doi.org/10.1007/BF01066988

Brannon, E. M., Wusthoff, C. J., Gallistel, C. R., & Gibbon, J. (2001). Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychological Science*, *12*(3), 238–243. https://doi.org/10.1111/1467-9280.00342

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cohen, L., & Dehaene, S. (1996). Cerebral networks for number processing: Evidence from a case of posterior callosal lesion. *Neurocase*, *2*(3), 155–174. https://doi.org/10.1080/13554799608402394

Cohen Kadosh, R., Lammertyn, J., & Izard, V. (2008). Are numbers special? An overview of chronometric, neuroimaging, developmental and comparative studies of magnitude representation. *Progress in Neurobiology*, *84*(2), 132–147.

Critchley, M. (1953). *The parietal lobes*. Oxford, England: Williams and Wilkins.

De Graaf, C., & Frijters, J. E. R. (1988). "Ratios" and "differences" in perceived sweetness intensity. *Perception & Psychophysics*, *44*(4), 357–362. https://doi.org/10.3758/BF03210417

Dehaene, S. (2001). Précis of The Number Sense. *Mind & Language*, *16*(1), 16–36. https://doi.org/10.1111/1468-0017.00154

Elmasian, R., & Birnbaum, M. H. (1984). A harmonious note on pitch: Scales of pitch derived from subtractive model of comparison agree with the musical scale. *Perception & Psychophysics, 36*(6), 531–537. https://doi.org/10.3758/BF03207513

Eisler, H. (1963). Magnitude scales, category scales, and Fechnerian integration. *Psychological Review, 70*(3), 243–253. https://doi.org/10.1037/h0043638

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf and Härtel

Fechner, G. T., Boring, E. G., Adler, H. E., & Howes, D. H. (1966). *Elemente der Psychophysik*. New York: Holt, Rinehart and Winston. (Original work published 1860).

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*(7), 307–314. https://doi.org/10.1016/j.tics.2004.05.002

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition, 44*(1–2), 43–74.

Garner, W. R. (1954). Context effects and the validity of loudness scales. *Journal of Experimental Psychology, 48*(3), 218–224. https://doi.org/10.1037/h0061514

Garner, W. R. (1954). A Technique and a Scale for Loudness Measurement. *The Journal of the Acoustical Society of America, 26*(1), 73–88. https://doi.org/10.1121/1.1907294

Gershman, S., Niv, Y. (2013). Perceptual estimation obeys Occam's razor. *Frontiers in Psychology, 4*, 623.

Gibson, E. J. (1969). *Principles of perceptual learning and development*. East Norwalk: Appleton-Century-Crofts.

Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2010). Non-symbolic arithmetic abilities and mathematics achievement in the first year of formal schooling. *Cognition, 115*(3), 394–406. https://doi.org/10.1016/j.cognition.2010.02.002

Grace, R., J. Morton, N., D. Ward, M., J. Wilson, A., & Kemp, S. (2018). Ratios and differences in perceptual comparison: A reexamination of Torgerson's conjecture. *Journal of Mathematical Psychology*, *85*, 62–75.

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(28), 11116–11120. https://doi.org/10.1073/pnas.1200196109

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668. https://doi.org/10.1038/nature07246

Hartley, A. A. (1977). Mental measurement in the magnitude estimation of length. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), 622–628. https://doi.org/10.1037/0096-1523.3.4.622

James, W. (1890). *The principles of psychology*. New York: Holt

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kraha, A., Turner, H., Nimon, K., Zientek, L., & Henson, R. (2012). Tools to Support Interpreting Multiple Regression in the Face of Multicollinearity. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00044

Krantz, D. H. (1972). A theory of magnitude estimation and cross-modality matching. *Journal of Mathematical Psychology*, *9*(2), 168–199. https://doi.org/10.1016/0022-2496(72)90025-9

Krueger, L. E. (1984). Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgments. *Perception & Psychophysics*, *35*(6), 536–542. https://doi.org/10.3758/BF03205949

Krueger, L. E. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences*, *12*(2), 251–267.

Laming, D. (1991). Reconciling Fechner and Stevens? *Behavioral and Brain Sciences*, *14*(1), 188–191. https://doi.org/10.1017/S0140525X0006605X

Lipton, J. S., & Spelke, E. S. (2003). Origins of Number Sense: Large-Number Discrimination in Human Infants. *Psychological Science*, *14*(5), 396–401. https://doi.org/10.1111/1467-9280.01453

Masin, S. C. (2012). Fundamental Measurement of Perceived Length and Perceived Area. *International Journal of Psychological Studies*, *4*. https://doi.org/10.5539/ijps.v4n3p23

Masin, S. C. (2013). On the ability to directly evaluate sensory ratios. *Attention, Perception, & Psychophysics*, *75*(1), 194–204. https://doi.org/10.3758/s13414-012-0382-0

Masin, S. C. (2014). Test of the Ratio Judgment Hypothesis. *The Journal of General Psychology*, *141*, 130–150. https://doi.org/10.1080/00221309.2014.883355

Masin, S. C., Brancaccio, A., & Tomassetti, A. (2019). Tests of the abilities to judge ratios of extensive and intensive sensory magnitudes. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-019-01710-x

McBride, R. L. (1983). A JND-scale/category-scale convergence in taste. *Perception & Psychophysics*, *34*(1), 77–83. https://doi.org/10.3758/BF03205899

Mcburney, D. H. (1966). Magnitude estimation of the taste of sodium chloride after adaptation to sodium chloride. *Journal of Experimental Psychology*, *72*(6), 869–873. https://doi.org/10.1037/h0023866

Mellers, B. A., Davis, D. M., & Birnbaum, M. H. (1984). Weight of evidence supports one operation for "ratios" and "differences" of heaviness. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(2), 216-230. https://doi.org/10.1037/0096-1523.10.2.216

Murofushi, K. (1997). Numerical matching behavior by a chimpanzee (*Pan troglodytes*): Subitizing and analogue magnitude estimation. *Japanese Psychological Research*, *39*, 140–153.

Murray, D. J. (1993). A perspective for viewing the history of psychophysics. *Behavioral and Brain Sciences*, *16*(1), 115–137. https://doi.org/10.1017/S0140525X00029277

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annual Review of Neuroscience*, *32*, 185–208. https://doi.org/10.1146/annurev.neuro.051508.135550

Norman, J. F., Adkins, O. C., & Pedersen, L. E. (2016). The visual perception of distance ratios in physical space. *Vision Research*, *123*, 1–7. https://doi.org/10.1016/j.visres.2016.03.009

Nys, J., & Content, A. (2012). Judgement of discrete and continuous quantity in adults: Number counts! *Quarterly Journal of Experimental Psychology (2006)*, *65*(4), 675–690. https://doi.org/10.1080/17470218.2011.619661

Parker, S., Schneider, B., & Kanow, G. (1975). Ratio scale measurement of the perceived lengths of lines. *Journal of Experimental Psychology. Human Perception and Performance*, *104*(2), 121–129.

Piazza, M. (2011). Neurocognitive Start-Up Tools for Symbolic Number Representations. *Space, Time and Number in the Brain*, 267–285.

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547–555. https://doi.org/10.1016/j.neuron.2004.10.014

Poulton, E. C. (1968). The new psychophysics: Six models for magnitude estimation. *Psychological Bulletin*, *69*(1), 1-19.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, *80*(1), 127–158. https://doi.org/10.1016/S0010-0277(00)00156-6

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science*, *19*(6), 607–614. https://doi.org/10.1111/j.1467-9280.2008.02130.x

Rule, S. J., Curtis, D. W., & Mullin, L. C. (1981). Subjective ratios and differences in perceived heaviness. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 459-466. https://doi.org/10.1037//0096-1523.7.2.459

Stein, J. F. (1992). The representation of egocentric space in the posterior parietal cortex. *Behavioral and Brain Sciences*, *15*(4), 691–700. https://doi.org/10.1017/S0140525X00072605

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Stevens, S. S. (1959). Cross-modality validation of subjective scales for loudness, vibration, and electric shock. *Journal of Experimental Psychology*, *57*(4), 201–209.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, *54*(6), 377–411.

Stevens, J. C., & Marks, L. E. (1965). Cross-modality matching of brightness and loudness. *Proceedings of the National Academy of Sciences of the United States of America*, *54*(2), 407–411.

Stevens, J. C., & Marks, L. E. (1980). Cross-modality matching functions generated by magnitude estimation. *Perception & Psychophysics, 27*(5), 379–389.

Schneider, B., Parker, S., Farrell, G., & Kanow, G. (1976). The perceptual basis of loudness ratio judgments. *Perception & Psychophysics*, *19*(4), 309–320. https://doi.org/10.3758/BF03204236

Starr, A., Libertus, M. E., & Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences of the United States of America, 110*(45), 18116–18120. https://doi.org/10.1073/pnas.1302751110

Teghtsoonian, R. (1971). On the exponents in Stevens' law and the constant in Ekman's law. *Psychological Review*, *78*(1), 71–80. https://doi.org/10.1037/h0030300

Tomonaga, K. (2002). Enumeration of briefly presented items by the chimpanzee (*Pan troglodytes*) and humans (*Homo sapiens*). *Animal Learning & Behavior. 30*(2), 143–157.

Torgerson, W. S. (1961). Distances and ratios in psychophysical scaling. *Acta Psychologica, 19*, 201-205.

Tosto, M. G., Petrill, S. A., Malykh, S., Malki, K., Haworth, C. M. A., Mazzocco, M. M. M., … Kovas, Y. (2017). Number Sense and Mathematics: Which, When and How? *Developmental Psychology*, *53*(10), 1924–1939. https://doi.org/10.1037/dev0000331

VanMarle, K. (2015). Foundations of the Formal Number Concept: How Preverbal Mechanisms Contribute to the Development of Cardinal Knowledge. In D. C. Geary, D. B. Berch, & K. M. Koepke (Eds.), *Mathematical Cognition and Learning*, 175–199.

Veit, C. T. (1978). Ratio and subtractive processes in psychophysical judgment. *Journal of Experimental Psychology: General*, *107*(1), 81-107. https://doi.org/10.1037/0096-3445.107.1.81

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488.

Walsh, V. (2003). Cognitive Neuroscience: Numerate Neurons. *Current Biology*, *13*(11), R447–R448. https://doi.org/10.1016/S0960-9822(03)00368-3

Webb, B., & Wystrach, A. (2016). Neural mechanisms of insect navigation. *Current Opinion in Insect Science*, *15*, 27–39. https://doi.org/10.1016/j.cois.2016.02.011

Zarate, J. M., Ritson, C. R., & Poeppel, D. (2012). Pitch-interval discrimination and musical expertise: Is the semitone a perceptual boundary? *The Journal of the Acoustical Society of America*, *132*(2), 984–993. https://doi.org/10.1121/1.4733535

Zeimbekis, J., & Raftopoulos, A. (2015). *The Cognitive Penetrability of Perception: New Philosophical Perspectives*. OUP Oxford.